БЕЛОРУССКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
Кафедра «Системы автоматизированного проектирования»

ЧИСЛЕННЫЕ МЕТОДЫ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Учебно-методическое пособие
для магистрантов специальности
1-31 81 12 «Прикладной компьютерный анализ данных»

Электронный учебный материал

Минск 2016

УДК: 517.63

Авторы:

В.М.Волков, И.Л. Ковалева

Рецензент:

кандидат физико-математических наук, доцент кафедры «Веб-технологии и компьютерное моделирование» механико-математического факультета БГУ М.В. Игнатенко.

В данном учебно-методическом пособии рассмотрены методы решения задачи Коши и краевых задач. Выполнена программная реализация описываемых методов в MATLAB, приведены результаты численных экспериментов. Каждый раздел содержит перечень конкретных упражнений, направленных на закрепление пройденного теоретического материала.

Особенностями пособия являются его обобщающий характер (изучение общей теории), сочетающийся с направленностью пособия на применение численных методов в решении практических задач.

Пособие предназначено для магистрантов, обучающихся по специальности «Прикладной компьютерный анализ данных».

Белорусский национальный технический университет
пр-т Независимости, 65, г. Минск, Республика Беларусь
Тел.(017) 293-95-67 факс (017) 292-71-53
E-mail: v.volkov@tut.by
il05kov@yahoo.com

Регистрационный № _БНТУ/ФИТР 50-69.2016

Оглавление

0.1. Введение

Строгость математических понятий и утверждений представляется наиболее привлекательным атрибутом математического аппарата с точки зрения его использования в качестве универсального языка науки. Описания закономерностей реального мира на языке математических абстракций и формулировка задачи в виде математической модели, учитывающей основные закономерности поведения объекта, позволяет получать новые знания об этом объекте исключительно математическими методами без необходимости непосредственного контакта с ним. Например, широкий круг явлений электродинамики может быть всесторонне изучен на основе анализа уравнений Максвелла. Аналогичным образом, глубокое теоретическое осмысление закономерностей гидродинамических течений доступно посредством анализа уравнений Навье–Стокса. Данный метод теоретических исследований, в основе которого математическая формулировка задачи и использование адекватного математического аппарата для ее решения, получил название математическое моделирование.

В последние пол века успехи прикладных математических методов в различных областях науки и техники во многом обязаны интенсивному использованию техники численного анализа. Благодаря численным методам математический язык стал еще более универсальным и обстоятельным. С другой стороны, развитие компьютерных технологий сделало этот язык более простым и доступным широкому кругу исследователей в различных областях знаний.

В настоящее время компьютерные технологии приобрели статус третьей силы в арсенале научных методов исследования, дополняя и расширяя возможности традиционных экспериментальных и теоретических подходов. Эффективность компьютерного моделирования достигается благодаря численным методам, отрывающим широчайшие возможности приближенного анализа математических моделей. Техника численного эксперимента ассоциируются с решением масштабных проблем, таких как прогноз погоды, расчет орбит космических аппаратов, аэродинамика сверхзвуковых скоростей, безопасность ядерных реакторов, оптимизация сложных технических систем и др., для которых другие подходы оказываются мало эффективными или вовсе неприменимы.

Предлагаемое учебное пособие содержит краткое изложение методов численного анализа начальных и краевых задач для обыкновенных дифференциальных уравнений. Пособие ориентировано на освещение практических вопросов решения дифференциальных задач, содержит большое число примеров и предназначено для студентов инженерных специальностей второй ступени обучения.

Глава 1

Численные методы решения задачи Коши

1.1. Постановка задачи

В общем случае дифференциальная задача может быть сформулирована для системы уравнений первого порядка

$$\frac{du_k}{dt} = f_k(t, u_1, u_2, \ldots, u_N), \quad k = 1, \ldots, N, \tag{1.1}$$

где $u_k = u_k(t)$ — искомые функции одной переменной, а f_k — заданные функции $N + 1$ переменных.

Примером дифференциальной задачи является система уравнений движения пружинного маятника :

$$\frac{dz}{dt} = v_z, \quad \frac{dv_z}{dt} = -\omega^2 z, \tag{1.2}$$

$$z(t_0) = z^0, \quad v_z(t_0) = v_z^0. \tag{1.3}$$

Здесь z и v_z — вертикальная координата и скорость груза, точка $z = 0$ соответствует положению равновесия, в которой сила упругости взаимно компенсируется с силой тяжести, постоянная ω — циклическая частота колебаний маятника,

$$\omega = \sqrt{\frac{k}{m}}$$

k — коэффициент упругости пружины m — масса груза, подвешенного на пружине.

Решение дифференциальной задачи не может быть однозначно определено уравнениями (1.2). Интуитивно понятно, что положение маятника в момент времени $t = t_1$ не может быть установлено с полной определенностью при отсутствии информации о его положении и скорости в некоторый предыдущий момент времени $t_0 < t_1$. В рассматриваемом примере такие дополнительные **начальные условия** заданы уравнениями (1.3).

Таким образом, для однозначного определения решения дифференциальной задачи система дифференциальных уравнений должна быть дополнена дополнительными условиями, например, в виде алгебраических уравнений, определяющих значения искомых функций в одной и той же или различных точках на оси независимой переменной. Такие дополнительные условия принято называть начальными условиями, когда значения искомых функций задается в одной точке, или краевыми условиями в противном случае.

Задачи с начальными условиями (**задачи Коши**) широко используются для моделирования динамических систем. Существование и единственность решения задачи Коши ассоциируется с принципом детерминизма. Как правило, если уравнения и начальные условия сформулированы корректно на физическом уровне строгости, то и с математической точки зрения задача также является корректной в смысле существования и единственности решения. Число дополнительных начальных условий, для однозначного определения обычно совпадает с количеством уравнений системы.

В дальнейшем мы будем использовать постановку задачи Коши для системы ОДУ первого порядка:

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(t, \mathbf{u}), \quad t \in [0, T], \tag{1.4}$$

$$\mathbf{u}(0) = \mathbf{u_0}, \tag{1.5}$$

где

$$\mathbf{u} = \mathbf{u}(t) = [u_1(t), u_2(t), \dots u_N(t)]^T,$$

$$\mathbf{f}(t, \mathbf{u}) = [f_1, f_2, \dots, f_N]^T,$$

$$f_k = f_k(t, u_1(t), u_2(t), \dots u_N(t)), \quad k = 1, \dots, N,$$

$$\mathbf{u}_0 = [u_1^0, u_2^0, \dots, u_N^0]^T.$$

В простейшем случае задача Коши включает одно уравнение и соответствующее начальное значение искомой функции при $t = 0$.

Упражнение 1.1.

1. Сформулируйте задачу Коши для уравнений (1.2), (1.3) в эквивалентном виде для одного дифференциального уравнения второго порядка.

2. Сформулируйте следующую задачу Коши

$$-\frac{du^3}{dt^3} + \frac{d^2u}{dt^2} + \frac{du}{dt} = f(t, u), \quad t \in [t_0, T], \tag{1.6}$$

$$\frac{d^2u}{dt^2}\,|_{t=t_0} = g_2, \quad \frac{du}{dt}\,|_{t=t_0} = g_1, \quad u(t)\,|_{t=t_0} = g_0, \tag{1.7}$$

в виде эквивалентной системы дифференциальных уравнений первого порядка.

1.2. Метод Эйлера. Явные и неявные схемы. Понятие устойчивости

Метод Эйлера — самый элементарный метод численного интегрирования задачи Коши для обыкновенных дифференциальных уравнений. Схема данного метода, например, может быть получена как следствие приближенного представления решения дифференциальной задачи ((1.4)), ((1.5)) отрезком степенного ряда :

$$u(t_0+\tau) = u(t_0) + \tau u'(t_0) + \frac{\tau^2}{2!}u''(t_0) + O(\tau^3). \tag{1.8}$$

При подстановке $u'(t_0) = f(t_0, u(t_0))$ в уравнение (1.8), пренебрегая членами высшего порядка малости, мы приходим к следующему уравнению для приближенного решения задачи (1.4),(1.5):

$$u(t_0+\tau) \simeq u(t_0) + \tau f(t_0, u(t_0)). \tag{1.9}$$

Таким образом, нам известно значение искомого решения задачи (1.4)–(1.5) при $t = t_0$, и мы можем вычислить приближенное значение искомого решения при $t = t_0 + \tau$ согласно (1.9). Применяя формулу (1.9), шаг за шагом, мы приходим к рекурсивному алгоритму численного интегрирования задачи Коши, (1.4)–(1.5), который позволяет вычислить приближенное решение для произвольного $t \in [0, T]$:

$$U(t_{k+1}) = U(t_k) + \tau f(t_k, U(t_k)), \quad t_k = k\tau, k = 0, 1, 2, \ldots. \tag{1.10}$$

Формула (1.10) известна как **метод Эйлера**.

Заметим, что решение в методе Эйлера вычисляется непосредственно по явной формуле. Данный класс алгоритмов принято называть **явными** .

Мы будем также называть метод Эйлера **одношаговым**, подчеркивая тем самым то, что для вычисления нового значения U_{k+1} при $t = t_{k+1}$ мы используем значение решения только в одной предыдущей точке при $t = t_k$ (на дистанции одного шага от искомого решения в новой точке).

Вычислительная сложность алгоритмов решения дифференциальных задач принято характеризовать количеством вычислений функций $f(t, u)$ на одном шаге численного интегрирования. Для метода Эйлера на каждом шаге требуется однократное вычисление $f(t, u)$.

Отличие точного решения $u(t_{k+1})$ и его приближенного значения (1.10) (при условии, что $u(t_k)$ определено точно) называется **локальной погрешностью** , т.е. погрешностью на одном шаге численного интегрирования:

$$\delta(t_k+\tau) = U(t_{k+1}) - u(t_{k+1}) = \frac{\tau^2}{2!}u''(t_k) + O(\tau^3). \tag{1.11}$$

Для определения глобальной погрешности при произвольном значении $t = T$ мы должны учитывать кумулятивный эффект накопления ошибки, обусловленный рекурсивной структурой алгоритма. Кроме того, необходимо убедиться, что алгоритм устойчив и локальная ошибка, полученная на одном шаге не испытывает неограниченного роста на последующих шагах.

Устойчивость численных методов для решения задачи Коши (1.4), (1.5) можно определить требованием выполнения оценки

$$|U_k| \leqslant C_1|U_0\| + C_2 \max_{0<m<k-1} |f(t_m, U_m)|, \tag{1.12}$$

где C_1 и C_2 — положительные постоянные, не зависящие от τ, $U_k = U(t_k)$.

Неравенство (1.12) означает, что приближенное решение непрерывно зависит от входных данных: малые возмущения начальных условий и правой приводят к ограниченным отклонениям траектории решения на произвольном отрезке $t \in [0, T]$.

Для исследования устойчивости методов численного решения задачи Коши, как правило, рассматривается стандартная тестовая задача:

$$\frac{du}{dt} = \lambda u, \quad \lambda < 0. \tag{1.13}$$

Заметим, что решение уравнения (1.13) при отрицательном значении λ является ограниченным и устойчивым:

$$u(t) = u(0)\exp(-|\lambda|t). \tag{1.14}$$

Приближенное решение данной задачи при использовании метода Эйлера имеет вид:

$$U_{k+1} = (1 - \tau|\lambda|)U_k = (1 - \tau|\lambda|)^k U_0. \tag{1.15}$$

Легко видеть, что решение (1.15) будет ограниченным и устойчивым при выполнении условия

$$|1 - \tau|\lambda|| \leqslant 1. \tag{1.16}$$

Последнее неравенство можно рассматривать как **условие устойчивости** метода Эйлера. Как следует из (1.16), метод Элера является устойчивым для тестовой задачи (1.13) при условии

$$\tau < 2/|\lambda|. \tag{1.17}$$

Численный метод, который устойчив при выполнении некоторых ограничений на шаги дискретизации, называется **условно устойчивым**. Если для устойчивости метода никаких ограничений на шаги сетки не требуется, то такой метод называют **безусловно устойчивым**.

В общем случае системы линейных дифференциальных уравнений вида

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u}, \quad \mathbf{u} \in R^N, \quad A \in R^{N\times N}, \tag{1.18}$$

условие устойчивости имеет вид, аналогичный (1.17), где вместо $|\lambda|$ следует использовать максимальное по модулю собственное значение или норму матрицы A.

В качестве примера безусловно устойчивого метода решения задачи Коши отметим так называемый **неявный метод Эйлера**:

$$U(t_{k+1}) = U(t_k) + \tau f(t_{k+1}, U(t_{k+1})), \quad t_k = k\tau, k = 0, 1, 2, \ldots. \tag{1.19}$$

где в отличие от (1.10) решение выражается неявно и может быть вычислено посредством решения соответствующего, вообще говоря нелинейного, уравнения или системы уравнений. Применяя метод (1.19) к решению тестовой задачи (1.13), имеем

$$U_{k+1} = (1+\tau\lambda)^{-1}U_k = (1+\tau\lambda)^{-k}U_0. \tag{1.20}$$

Очевидно, что для любого $\tau > 0$ и $\lambda > 0$:

$$|U_{k+1}| \leqslant |U_0|, \tag{1.21}$$

и, согласно (1.12), неявный метод Эйлера является безусловно устойчивым.

Будем говорить, что численный метод сходится и имеет n-й **порядок точности** (**скорость сходимости** $O(\tau^n)$), если глобальная погрешность данного метода убывает пропорционально τ^n: $\delta_k = O(\tau^n)$.

Несложно показать, что глобальная погрешность метода Эйлера стремится к нулю при $\tau \to 0$:

$$|\delta_k| \leqslant C\tau, \quad k = 1, 2, \ldots. \tag{1.22}$$

Здесь C — постоянная, не зависящая от τ.

Как следует из оценки (1.22), метод Эйлера (1.4)–(1.5) сходится, имеет первый порядок точности и его погрешность убывает пропорционально τ. Кумулятивный эффект накопления ошибки приводит к тому, что глобальная ошибка на порядок превосходит локальную погрешность. Так, если локальная погрешность имеет порядок $O(\tau^p)$, то глобальная погрешность при условии устойчивости алгоритма, как правило, ограничена величиной $O(\tau^{p-1})$. В силу низкой скорости сходимости и условной устойчивости явный метод Эйлера имеет низкую эффективность и практически не используются для решения реальных задач.

Пример 1.1. Рассмотрим пример, демонстрирующий типичные проявления эффекта потери устойчивости, когда условия устойчивости оказываются нарушенными. Для этого рассмотрим явный метод Эйлера применительно к модельному уравнению вида (1.13), $\lambda = 1.5$ с начальными условиями $u(0) = 2$. Программная реализация и результаты численных экспериментов представлены ниже. Несложно видеть, что при $\tau < \tau_0/2$, где τ_0 опрелено условием устойчивости (1.17), приближенное решение мало отличается от точного. В случае $\tau_0/2 < \tau < \tau_0$ приближенное решение теряет свойство монотонности, но остается асимптотически ограниченным и устойчивым. Наконец, когда $\tau > \tau_0$ метод перестает сходится из-за потери устойчивости. Продолжая вычисления, легко убедиться, что неустойчивость приводит к неограниченному росту решения, в то время как точное решение стремится к нулю при $t \to 0$.

```
%%% Устойчивость & Сходимость
%%% метод Эйлера для уравнения u'=-lambda*u
lambda = 1.5;
T = 10.
U0 = 2;
tau_0 = 2/lambda;
NN = round(T/tau_0)*10;
u = zeros(3,NN);
t = u;
n = 0;
t(:,1) = 0;
```

```
u(:,1) = U0;
for tau = [0.2, 0.8 1.1]*tau_0
n = n+1;
N = T/tau;
U = U0;
for m = 1:N-1
U = (1-tau*lambda)*U;
u(n,m+1) = U;
t(n,m+1) = tau*m;
end
NN(n) = N;
end
plot(t(1,1:NN(1)),u(1,1:NN(1)),'.-',...
t(2,1:NN(2)),u(2,1:NN(2)),'o-',...
t(3,1:NN(3)),u(3,1:NN(3)),'o-')
legend('\tau=0.2*\tau_0',...
'\tau=0.8*\tau_0','\tau=1.1*\tau_0')
xlabel('t')
ylabel('U(t)')
grid
```

Типичная картина развития неустойчивости представлена на Рис. 1.1. Аналогичное поведение решения характерно для большинства явных численных методов при использовании шага дискретизации не удовлетворяющего условиям устойчивости. Отметим, что существуют также абсолютно неустойчивые методы, сохраняющие подобное поведение при произвольном, сколь угодно малом значении шага численного интегрирования.

Упражнение 1.2.

1. Исследовать устойчивость двухстадийного численного метода решения задачи Коши для уравнения (1.13), состоящего в чередовании явной и неявной формул Эйлера (1.10) и (1.19) на нечетных и четных шагах соответственно.

2. Оцените локальную погрешность неявного метода Эйлера (1.19) и метода, рассмотренного в предыдущем упражнении.

3.Пусть задачи Коши была решена трижды с различными значениями шага численного интегрирования: $\tau = \tau_0$; $\tau = \tau_0/2$; и $\tau = \tau_0/4$, где τ_0 достаточно мало. Оцените погрешность метода и его порядок точности, используя упомянутые выше три приближенные решения, полагая асимптотическое поведение погрешности $|\delta| \simeq C\tau^p$ и считая точное решение неизвестным.

4. Воспроизведите численный эксперимент предыдущего пункта 3, используя численный алгоритм, рассмотренные в примере выше, используя $\tau_0 = 0.01$. Оцените фактическую погрешность численного метода, используя точное решение задачи и сравните результат с оценкой (1.22), а также с оценкой, полученной в предыдущем упражнении 3.

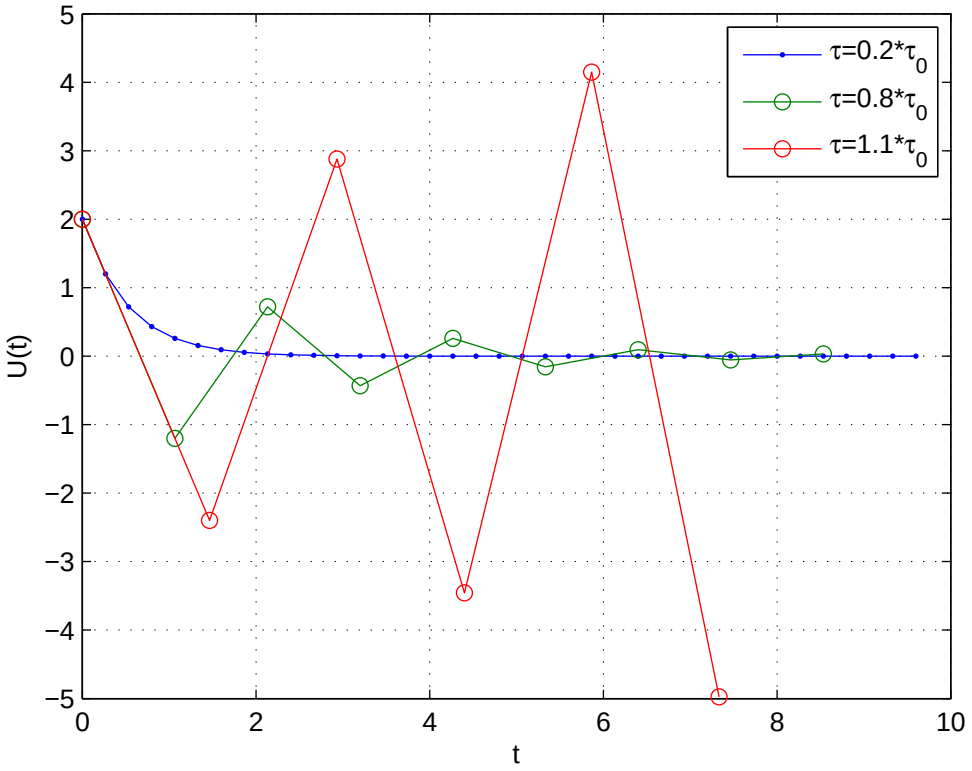

Рис. 1.1. Потеря устойчивости явного метода Эйлера при нарушении условия устойчивости (1.17)

1.3. Методы Рунге – Кутты

Методы Рунге – Кутты представляют собой семейство методов численного анализа задач Коши для систем обыкновенных дифференциальных уравнений вида (1.4)–(1.5). Наибольшей распространение в практике вычислений получили явные одношаговые **методы Рунге – Кутты** следующего вида:

$$U_{k+1} = U_k + \sum_{m=1}^{s} b_m K_m, \tag{1.23}$$

где s — целая постоянная, определяющая количество стадий вычислений, связанных с пересчетом функции правой части, в пределах одного шага численного интегрирования,

$$\begin{aligned} &K_1 = \tau f(t_k,\, U_k),\\ &K_2 = \tau f(t_k + c_2\tau,\, U_k + a_{21}K_1),\\ &\ldots\ldots\ldots\ldots\\ &K_s = \tau f(t_k + c_s\tau,\, U_k + \sum_{i=1}^{s-1} a_{si}K_i). \end{aligned} \tag{1.24}$$

Коэффициенты b_m, c_m и a_{km} определяются из условия максимального порядка малости локальной погрешности для заданного числа стадий. Для наглядности коэффициенты метода удобно представлять в виде **таблицы Бутчера** (в честь Дж. Бутчера (John C. Butcher)):

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cccccc} 0 & 0 & & & & & \\ c_2 & a_{21} & 0 & & & & \\ \vdots & \vdots & \ddots & & & & \\ c_s & a_{s1} & a_{s2} & \ldots & a_{ss-1} & 0 & \\ \hline & b_1 & b_2 & \ldots & \ldots & b_s & \end{array} \tag{1.25}$$

Примечательно, что условие максимального порядка малости локальной погрешности не обеспечивает однозначности определения коэффициентов метода Рунге – Кутты. Пример коэффициентов наиболее популярного метода Рунге – Кутты четвертого порядка точности представлен ниже в таблице:

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{1.26}$$

Несложно заметить, что для $s = 1$ одностадийный метод Рунге – Кутты полностью совпадает с методом Эйлера. Следует отметить, что порядок точности методов Рунге – Кутты (1.23)–(1.24) при $1 \leqslant s \leqslant 4$ совпадает с количеством стадий. Для более сложных схем, $s > 4$, порядок точности p ($|\delta_k| = |U_k - u(t_k)| = O(\tau^p)$)

не превосходит s и увеличивается с увеличением количества стадий как показано в таблице ниже.

s	1	2	3	4	5	6	7	8	9	10
p	1	2	3	4	4	5	6	6	7	7

(1.27)

Вычислительная сложность метода Рунге – Кутты зависит от количества вычислений функции $f(t,u)$ и растет пропорционально числу стадий s, поскольку на каждой стадии значение функции вычисляется один раз. Учитывая не всегда пропорциональную зависимость порядка точности от числа стадий, можно сделать вывод, что наиболее интенсивный рост эффективности метода Рунге – Кутты происходит до четвертого порядка включительно, после чего вычислительная сложность растет быстрее, нежели порядок точности (см.таблицу 1.27) . Этим, в частности, можно объяснить наибольшую популярность среди методов Рунге – Кутты именно четырех-этапной версии, обеспечивающей соответственно четвертый порядок точности.

Информация относительно порядка точности численного метода может быть использована для апостериорной оценки фактической погрешности приближенного решения. В частности, скорость сходимости определяет асимптотическое поведение погрешности решения при $\tau \to 0$:

$$|\delta(t_k)| = O(\tau^p) \;\Rightarrow \delta(t_k) \simeq C\tau^p, \tag{1.28}$$

где C — постоянная не зависящая от τ.

Имея два приближенных решения $U_{(1)}(t_k)$ и $U_{(2)}(t_{2k})$, полученные при использовании различных шагов интегрирования, например, $\tau_1 = \tau_0$ и $\tau_2 = \tau_0/2$, мы можем выразить приближенно погрешность следующим образом:

$$U_{(1)}(t_k) \simeq u(t_k) + C\tau_0^p, \quad U_{(2)}(t_{2k}) \simeq u(t_{2k}) + 2^{-p}C\tau_0^{-p},$$

$$\delta_{(2)}(t_{2k}) \simeq \frac{U_{(1)}(t_k) - U_{(2)}(t_{2k})}{2^p - 1}. \tag{1.29}$$

Уравнение (1.29) при достаточно малом значении τ_0 дает реалистичную оценку погрешности приближенного решения при $\tau = \tau_0/2$.

Упражнение 1.3.

1. Докажите, что приближенное решение метода Рунге–Кутты четвертого порядка точности (1.25), (1.26) для модельной задачи

$$\frac{du}{dt} = \lambda u, \quad u(0) = u_0, \tag{1.30}$$

может быть представлено в следующем виде:

$$U(t_k) = u_0 \left(1 + \tau\lambda + \frac{(\tau\lambda)^2}{2!} + \frac{(\tau\lambda)^3}{3!} + \frac{(\tau\lambda)^4}{4!}\right)^k, \tag{1.31}$$

где выражение в скобках есть отрезок степенного ряда разложения функции $\exp(k\tau\lambda)$.

2. Оцените величину погрешности метода Рунге – Кутты четвертого порядка точности используя формулы (1.31) и (1.29) при $\tau_0 = 0.02$, $\tau_0 = 0.01$, $\tau_0 = 0.005$ применительно к тестовой задаче (1.30) при $\lambda = 1$, $t = 1$. Сравните полученную оценку погрешности с фактической погрешностью $\delta(1) = U(1) - u(1)$, используя точное решение тестовой задачи $u(t) = u_0 \exp(\lambda t)$.

3. Построить шрафик функции

$$W_{RK}(s) = \left| 1 + s + \frac{s^2}{2!} + \frac{s^3}{3!} + \frac{s^4}{4!} \right|, \tag{1.32}$$

которая может быть использована в качестве модуля функции передачи для расчета приближенного решения методом Рунге - Кутты четвертого порядка точности на одном шаге, применительно к решению уравнения (1.30), $s = \tau\lambda$. Используя утверждение упражнения 1., определите область устойчивости метода Рунге – Кутты четвертого порядка точности, т.е. область, где $|W_{RK}(s)| \leqslant 1$ при $\lambda < 0$.

4 Сравните функцию передачи (1.32), и функцию передачи метода Эйлера (1.15), модуль которой

$$|W_E(s)| = |1 + s| \,. \tag{1.33}$$

Какой из методов обеспечивают большую область устойчивости и позволяет использовать больший шаг τ.

1.4. Многошаговые методы

Мы рассмотрели семейство методов Рунге – Кутты, относящихся к классу одношаговых явных методов. Гибкость в использовании и достаточно высокая эффективность применительно к широкому кругу дифференциальных задач является бесспорным достоинством данного семейства методов. Тем не менее, увеличение скорости сходимости выше четвертого порядка в методах Рунге – Кутты сопряжено с некоторым дополнительным ростом вычислительных затрат (см. зависимость порядка точности от числа стадий в методе Рунге – Кутты (1.27)). Эффективность одношаговых методов Рунге – Кутты может существенно ухудшится, если вычисление функции правой части системы дифференциальных уравнений требуе больших вычислительных затрат. Такого рода недостатки можно преодолеть, используя многошаговый подход к построению дискретной модели дифференциальной задачи.

Линейные **многошаговые методы** для решения задачи Коши (1.4)–(1.5) имеет следующий общий вид:

$$U_k + a_1 U_{k-1} + \ldots + a_m U_{k-m} = \tau \left[b_0 f_k + b_1 f_{k-1} + \ldots + b_m f_{k-m} \right], \tag{1.34}$$

где $f_k = f(t_k, U_k)$, $U_k = U(t_k)$, а значения коэффициентов a_n и b_n определяются из условий минимума локальной погрешности метода. Максимальный порядок точности, который формально может быть получен в рамках m-шаговой схемы (1.34),

достигает $O(\tau^{2m})$. Тем не менее, многошаговые схемы максимального порядка точности оказываются абсолютно неустойчивыми и непригодными для практического использования. Приемлемые условия устойчивости могут быть достигнуты лишь для многошаговых методов с порядком точности $O(\tau^m)$ при использовании явной схемы метода, или на один (два) порядка выше для неявных методов с нечетным (четным) значением числа шагов m.

В качестве многошаговых методов, способных обеспечивать устойчивость, можно отметить семейство **методов Адамса**,

$$U_k - U_{k-1} = \tau \left[b_0 f_k + b_1 f_{k-1} + \ldots + b_m f_{k-m} \right]. \tag{1.35}$$

среди которых явные методы **Адамса–Башфорта** , для которых $b_0 = 0$, и неявные **методы Адамса – Моултона** ($b_0 \neq 0$). Максимальный порядок точности m-шагового метода Адамса – Башфорта и Адамса– Моултона достигает $O(\tau^m)$ и $O(\tau^{m+1})$ соответственно. Простейшим примером метода Адамса–Башфорта может служить одношаговый явный метод Эйлера ($b_0 = 0,\ b_1 = 1$).

Методы Адамса – Башфорта более высокого порядка точности с оценками главного члена локальной ошибки δ_u имеют вид:

$$U_k = U_{k-1} + \tau V_k, \tag{1.36}$$

где

$$\begin{aligned}
V_k &= \left[\frac{3}{2} f_{k-1} - \frac{1}{2} f_{k-2} \right], && \delta_u = \frac{5\tau^3}{12} u''', \\
V_k &= \left[\frac{23}{12} f_{k-1} - \frac{16}{12} f_{k-2} + \frac{5}{12} f_{k-3} \right], && \delta_u = \frac{9\tau^4}{24} u^{(4)}, \\
V_k &= \left[\frac{55}{24} f_{k-1} - \frac{59}{24} f_{k-2} + \frac{37}{24} f_{k-3} - \frac{9}{24} f_{k-4} \right], && \delta_u = \frac{251\tau^5}{720} u^{(5)}, \\
V_k &= \left[\frac{1901}{720} f_{k-1} - \frac{2774}{720} f_{k-2} + \frac{2616}{720} f_{k-3} - \frac{1274}{720} f_{k-4} + \frac{251}{720} f_{k-5} \right], && \delta_u = \frac{95\tau^6}{2888} u^{(6)}.
\end{aligned} \tag{1.37}$$

Глобальна ошибка этих методов на порядок больше и варьируется от второго до пятого.

Методы Адамса– Моултона различного порядка точности также определяются общей формулой (1.36), где:

$$V_k = \tau\left[\frac{1}{2}f_k + \frac{1}{2}f_{k-1}\right], \qquad \delta_u = -\frac{\tau^3}{12}u''',$$

$$V_k = \tau\left[\frac{5}{12}f_k + \frac{8}{12}f_{k-1} - \frac{1}{12}f_{k-2}\right], \qquad \delta_u = -\frac{\tau^4}{24}u^{(4)},$$

$$V_k = \tau\left[\frac{9}{24}f_k + \frac{19}{24}f_{k-1} - \frac{5}{24}f_{k-2} + \frac{1}{24}f_{k-3}\right], \qquad \delta_u = -\frac{19\tau^5}{720}u^{(5)},$$

$$V_k = \tau\left[\frac{251}{720}f_k + \frac{646}{720}f_{k-1} - \frac{264}{720}f_{k-2} + \frac{106}{720}f_{k-3} - \frac{19}{720}f_{k-4}\right], \quad \delta_u = -\frac{3\tau^6}{160}u^{(6)}. \tag{1.38}$$

Несмотря на существенное сходство методов Адамса – Моултона и Адамса–Башфорта, принципиальное их отличие состоит в том, что методы Адамса – Моултона являются неявными и их решение не может быть выражено в явном виде, как в методе Адамса–Башфорта. В общем случае реализация неявных методов (1.38) приводит к решению на каждом шаге системы нелинейных алгебраических уравнений. Для этих целей обычно используются итерационные методы Ньютона или Пикара. Эффективным также представляется комбинация явных и неявных методов Адамса в схеме получившей название **предиктор – корректор**. Схема предиктор - корректор может рассматриваться как одна итерация в неявном методе, начальное приближение для которой вычисляется по явной схеме соответствующего порядка точности.

В качестве примера рассмотрим схему предиктор – корректор, построенную на трехшаговых методах Адамса. Пусть нам известно решение задачи в точках сетки t_{k-1}, t_{k-2}, t_{k-3}. Вычисляем неизвестное значение U_k в два этапа. На первом этапе, **предиктор**, мы используем трехшаговый явный метод Адамса–Башфорта, рассматривая полученное значение U_k как промежуточный результат:

$$\tilde{U}_k = U_{k-1} + h\left[\frac{23}{12}f_{k-1} - \frac{16}{12}f_{k-2} + \frac{5}{12}f_{k-3}\right]. \tag{1.39}$$

Затем мы корректируем полученное значение, используя для этого неявную формулу Адамса – Моултона:

$$U_k = U_{k-1} + h\left[\frac{5}{12}\tilde{f}_k + \frac{8}{12}f_{k-1} - \frac{1}{12}f_{k-2}\right], \tag{1.40}$$

где для вычисления $\tilde{f}_k = f(t_k, \tilde{U}_k)$. используем значение $\tilde{U}_k$, вычисленное на стадии предиктора. Стадия корректора (1.40) позволяет не только повысить точность решения, но и улучшить устойчивость метода. Примечательная деталь относительно методов Адамса состоит в том, что локальная погрешность для явных и неявных схем имеет противоположные знаки. (см. (1.37) и (1.38)). Кроме того, абсолютная величина локальной ошибки для неявных методов (1.38)) во много раз меньше по сравнению с явными (1.37).

Основное преимущество методов Адамса состоит в том, что, в отличие от методов Рунге – Кутты, функция правой части задачи на каждом шаге вычисляется всего один раз, независимо от порядка точности метода. Как следствие, вычислительная сложность многошаговых методов практически не зависит от порядка точности, что выгодно отличает их от одношаговых аналогов. Так, например, в методе Рунге – Кутты четвертого порядка точности функция правой части вычисляется четыре раза на каждом шаге, в то время как в явном методе Адамса можно ограничиться однократным вычислением практически для произвольного порядка точности. Таким образом, преимущества многошаговых методов становится более существенным при использовании формул более высокого порядка точности и особенно в случаях, когда вычисление функции правой части задачи (1.4)–(1.5) сопряжено со значительными вычислительными затратами.

Тем не менее, следует отметить также то, что для вычисления решения U_k в точке $t = t_k$ при использовании m-шагового метода нам необходимо знать решение в точках, $U_{k-1}, \ldots, U_{k-m}$. В начале вычислений значение U_{k-m} при $t = t_{k-m} = t_0$ определено начальными условиями. Для расчета остальных недостающих значений требуется использовать другие, например одношаговые методы типа методов Рунге – Кутты соответствующего порядка точности. Аналогичная ситуация может возникнуть при изменении шага τ в многошаговой схеме.

Для сравнения устойчивости явных и неявных многошаговых методов рассмотрим тестовую задачу

$$\frac{du}{dt} = \lambda u, \tag{1.41}$$

где λ — произвольное комплексное число. Решение задачи (1.41) экспоненциально устойчиво при $Re(\lambda) < 0$, т.е. область устойчивости дифференциальной задачи занимает левую полуплоскость комплексной плоскости. Для исследования устойчивости представим решение методов Адамса в степенном виде:

$$U_k = q^k. \tag{1.42}$$

Подстановка данного решения в формулы метода Адамса приводит к характеристическим уравнениям для величины q. Например, применение метода Адамса – Моултона второго порядка точности к решению задачи (1.42) приводит к следующему характеристическому уравнению:

$$q^k - q^{k-1} = \frac{\lambda\tau}{2}(q^k + q^{k-1}), \tag{1.43}$$

Уравнение (1.43) имеет корень

$$q = \frac{1 + \tau\lambda/2}{1 - \tau\lambda/2}. \tag{1.44}$$

Очевидно, что в области устойчивости дифференциальной задачи ($Re(\lambda) < 0$) мы имеем $|q| \leqslant 1$, и приближенное решение (1.42) также устойчиво, подобно решению дифференциальной задачи. В общем случае для оценки устойчивости используются так называемое **условие корней**: если все корни характеристического уравнения (полинома) локализованы внутри единичного круга комплексной плоскости, $|q| \leqslant 1$,

и отсутствуют кратные корни $|q| = 1$, то соответствующий многошаговый метод является устойчивым. Если условие корней не выполнено для $\lambda = 0$, то метод считается абсолютно неустойчивым. Таким образом, исследование устойчивости многошаговых методов сводится к проверке условия корней. Очевидно, что для рассмотренной тестовой задачи корни уравнения зависят от величины $\mu = \tau\lambda$. Множество точек комплексной плоскости $\mu \in C$, в которых выполняется условие корней называется **областью устойчивости метода**.

Если область устойчивости метода совпадает с областью устойчивости дифференциальной задачи, то такой метод называется A - **устойчивым**, или **абсолютно устойчивым**. К сожалению, не существует явных A - устойчивых методов. Более того, среди неявных линейных методов не существует методов выше второго порядка точности.

Если многошаговый метод не является A - устойчивым, то его область устойчивости ограничена и граница области устойчивости определяется соотношением между шагом τ и λ, вытекающим из характеристического уравнения при $|q| = 1$. Простейший способ определить границу области устойчивости состоит в том, чтобы нарисовать на комплексной плоскости множество точек $\mu = \tau\lambda$ для которых $|q| \equiv 1$. Например, в случае метода Адамса – Моултона третьего порядка точности характеристическое уравнение имеет вид

$$q - 1 = \tau\lambda \left[\frac{5}{12} q^2 + \frac{8}{12} q - \frac{1}{12} \right], \tag{1.45}$$

и область устойчивости определяется уравнением:

$$\mu = \tau\lambda = (q-1) \left[\frac{5}{12} q^2 + \frac{8}{12} q - \frac{1}{12} \right]^{-1}, \quad q = \exp(i\varphi), \quad \varphi \in [1, 2\pi]. \tag{1.46}$$

Области устойчивости для некоторых явных и неявных методов Адамса представлены на рис. 1.2. Несложно заметить, что область устойчивости неявных методов существенно превосходит область устойчивости явных аналогов, имеющих тот же порядок точности. С ростом порядка точности область устойчивости уменьшается.

Пример 1.2. В примере, рассмотренном ниже, мы сравним эффективность методов Рунге – Кутты и многошаговых методов Адамса, оценивая вычислительные затраты для получения заданной точности. В качестве тестовой задачи мы рассмотрим уравнение (1.2), для которого известно точное решение. Сравним эффективность методов четвертого порядка точности. Программная реализация алгоритмов и результаты численных экспериментов представлены ниже (см. рис. 1.2 и 1.3).

```
%Эффекктивность методов решения ОДУ
% Метод Рунге -- Кутты 4-го порядка (RC)
% Метод Адамса -- Башфорта 4-го порядка (AB)
% Метод предиктор -- корректор 4-го порядка (PC)
% Задача:   u''=-3u, t in [1,10];
% начальные условия: u(0)=0, u'(0)=3;
% точное решение: u(t)=sin(3t);
```

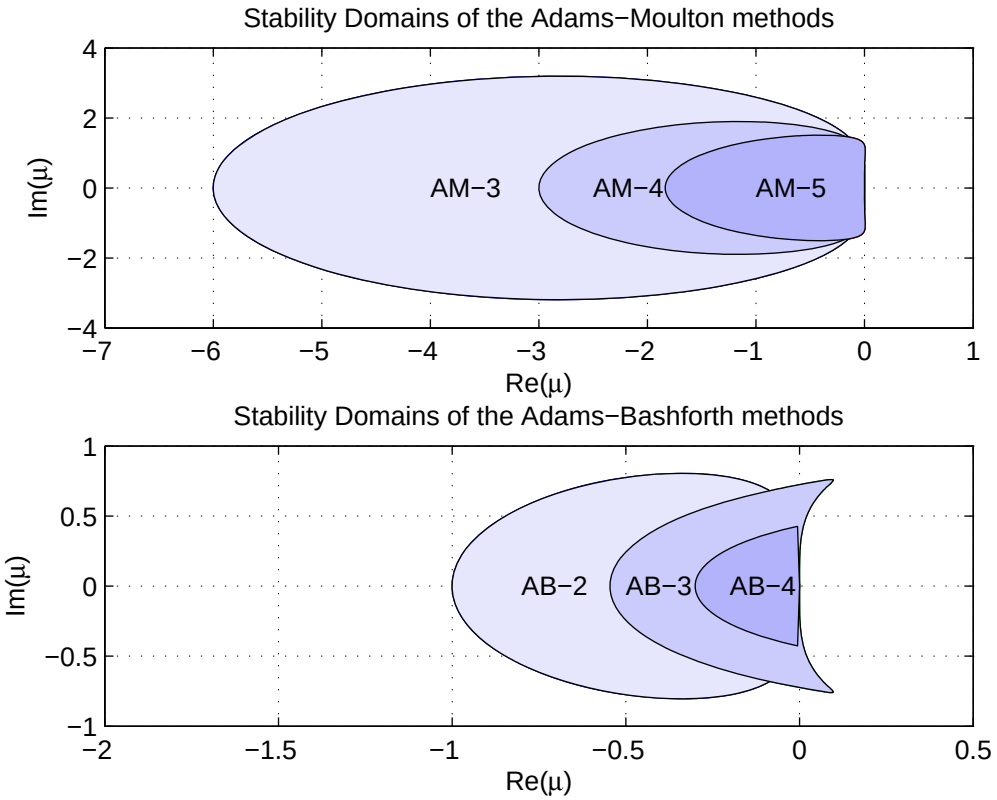

Рис. 1.2. Области устойчивости многошаговых явных и неявных методов Адамса.

```
f = @(t,u)[u(2); -9*u(1)];
T = 10;
tau = 0.01;
u = [0;3];
N = T/tau;
ts = 0:tau:T;
U = zeros(2,N+1);
%Метод Рунге -- Кутты 4 порядка
tic
U(:,1) = u;
for m = 1:N
t = (m-1)*tau;
K1 = tau*f(t,u);
K2 = tau*f(t+tau/2,u+K1/2);
K3 = tau*f(t+tau/2,u+K2/2);
K4 = tau*f(t+tau,u+K3);
u = u+(K1+2*K2+2*K3+K4)/6;
```

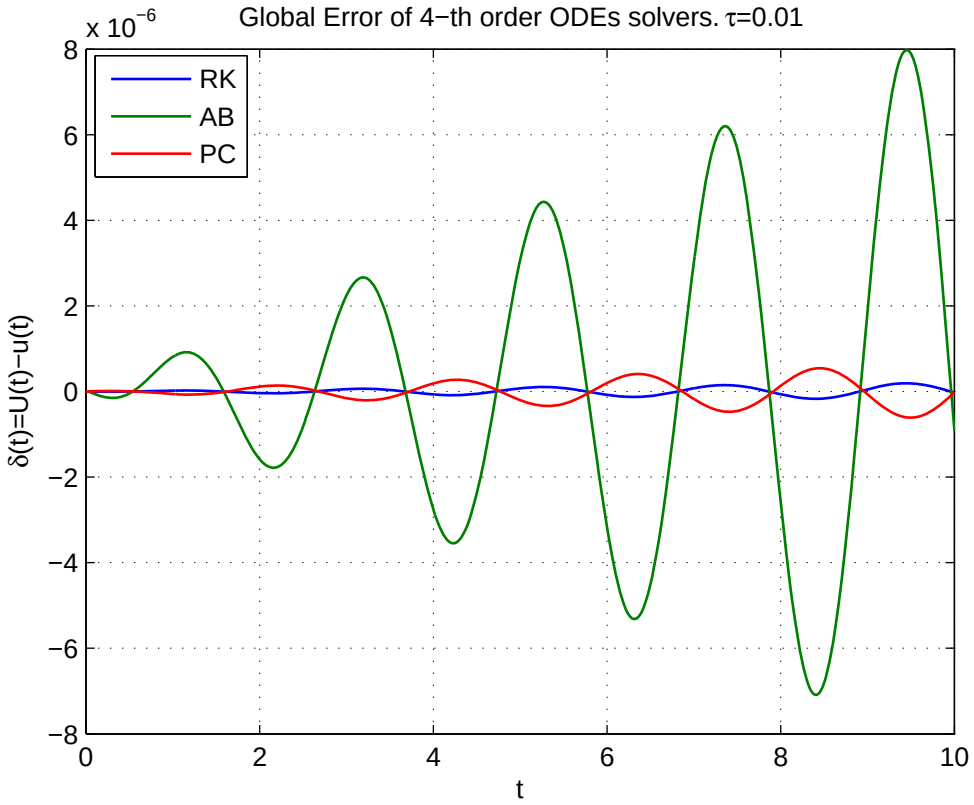

Рис. 1.3. Динамика погрешности методов четвертого порядка точности Рунге – Кутты (RK), Адамса – Башфорта (AB) предиктор – корректор (PC)

```
U(:,m+1) = u;
end
RK_time = toc
B = tau*[-9/24;37/24;-59/24;55/24];
F = zeros(2,N+1);
u = [0;3];
U1 = zeros(2,N+1);
U1(:,1) = u;
F(:,1) = f(0,u);
% Метод Адамса -- Башфорта 4 порядка
tic
for m = 1:3
t = (m-1)*tau;
K1 = tau*f(t,u);
K2 = tau*f(t+tau/2,u+K1/2);
K3 = tau*f(t+tau/2,u+K2/2);
```

```
K4 = tau*f(t+tau,u+K3);
u=u+(K1+2*K2+2*K3+K4)/6;
U1(:,m+1) = u;
F(:,m+1) = f(t+tau,u);
end
for m = 4:N
u = u+F(:,m-3:m)*B;
t = m*tau;
F(:,m+1) = f(t,u);
U1(:,m+1) = u;
end
AB_time = toc
%Predictor-Corrector on the base of 4-th order
% Метод предиктор -- корректор 4-го порядка
A = tau*[1/24;-5/24;19/24;9/24;];
u2 = [0;3];
U2 = zeros(2,N+1);
U2(:,1) = u2;
F(:,1) = f(0,u2);
tic
for m=1:3
t = (m-1)*tau;
K1 = tau*f(t,u2);
K2 = tau*f(t+tau/2,u2+K1/2);
K3 = tau*f(t+tau/2,u2+K2/2);
K4 = tau*f(t+tau,u2+K3);
u2 = u2+(K1+2*K2+2*K3+K4)/6;
U2(:,m+1) = u2;
F(:,m+1) = f(t+tau,u2);
end
for m = 4:N
u22 = u2+F(:,m-3:m)*B;%Predictor
t = m*tau;
F(:,m+1) = f(t,u22);
u2 = u2+F(:,m-2:m+1)*A;%Corrector
F(:,m+1) = f(t,u2);
U2(:,m+1) = u2;
end
PC_time = toc
y = sin(3*ts); % Exact solution;
plot(ts,(U(1,:)-y),ts,U1(1,:)-y,ts,U2(1,:)-y,'LineWidth',1)
title(['Global Error of 4-th order ODEs solvers.
\tau=',num2str(tau,3)])
xlabel('t')
ylabel('\delta(t)=U(t)-u(t)')
legend('RK','AB','PC');
```

```
grid on
```

Когда вычисления заканчиваются, в командном окне можно видеть время решения задачи для каждого из рассмотренных методов, Рунге – Кутты (RK), Адамса – Башфорта (AB) предиктор – корректор (PC):

```
RK_time =  0.1174
AB_time =  0.0518
PC_time =  0.0776
```

Несложно заметить, что ни один из рассмотренных методов не имеет бесспорного преимущества. Явный метод Адамса оказывается несколько быстрее своих конкурентов, но он проигрывает им в точности. Погрешности метода Адамса – Башфорта примерно в десять раз превосходит погрешность метода Адамса – Моултона, как это предсказывают оценки их локальных ошибок (см. (1.37), (1.38)).

Упражнение 1.4.

1. Внося необходимые коррективы в рассмотренном выше примере, замените программную реализацию методов Рунге – Кутты и Адамса – Башфорта на соответствующие стандартные функции MATLAB **ode45** и **ode15s**, в которых реализованы указанные методы. Сравните эффективность стандартной реализации рассмотренных методов, сопоставляя вычислительные затраты для достижения заданной точности.

2. Используя тестовую задачу, рассмотренную в примере выше и упражнении 1., проверьте, как значения входного параметра **RelTol** в функциях **ode45** и **ode15s** влияют на фактическую погрешностью численного решения.

3. Сравните эффективность решения задачи Коши с помощью MATLAB функций **ode45** и **ode15s** применительно к модифицированной тестовой задаче:

$$\begin{cases} \dfrac{du_1}{dt} = u_2, \\ \dfrac{du_2}{dt} = -\omega^2 \sin(u_1) \end{cases} \tag{1.47}$$

$$u_1(0) = 3, \quad u_2(0) = 0, \quad \omega = 3. \tag{1.48}$$

4. Исследовать устойчивость и локальную погрешность следующих многошаговых методов:

$$U_{k+1} = U_{k-1} + 2\tau f_k, \tag{1.49}$$

$$U_{k+1} = U_{k-1} + \frac{\tau}{2}\left(f_{k+1} + f_k + f_{k-1}\right), \tag{1.50}$$

$$U_{k+1} = U_k + \frac{\tau}{6}\left(6f_k - 3f_{k-1} + 3f_{k-2}\right), \tag{1.51}$$

1.5. Жесткие системы. Метод Гира и неявные методы Рунге – Кутты

Большинство методов, используемых для решения задачи Коши являются условно устойчивыми. Условия устойчивости налагают определенные ограничения на размер шага численного интегрирования. Например, чтобы обеспечить устойчивость метода Адамса – Моултона третьего порядка точности при решении тестовой задачи (1.41) в случае $\lambda \leqslant 0$ необходимо выполнение условия $\tau_0 \leqslant 6/|\lambda|$, что непосредственно определяется границей области устойчивости, представленной на рис. 1.2. В то же время для устойчивости явного метода Адамса – Моултона третьего порядка точности требуется размер шага в двенадцать раз меньше, чем для соответствующего неявного аналога.

Как правило, при решения задачи Коши для одного уравнения ограничения на размер шага, вытекающие из условий устойчивости, не являются более жесткими, нежели ограничения, продиктованные требованиями точности. Как следствие, при численном интегрировании одного уравнения явные методы представляются более предпочтительными. Однако в случае систем ОДУ может возникать иная ситуация. Например, рассмотрим следующую систему дифференциальных уравнений:

$$\frac{du}{dt} = -Au, \tag{1.52}$$

где, для большей наглядности, A — симметричная, положительно определенная матрица 2×2, собственные значения которой $\lambda_1 = 1e3$, $\lambda_2 = 1e-3$. С помощью преобразований подобия система уравнений (1.52) может быть приведена к диагональному виду

$$\frac{d\psi}{dt} = -D\psi, \quad D = P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad u = P\psi, \tag{1.53}$$

где P — матрица преобразования подобия. Сделанные преобразования показывают, что решение задачи представлено двумя затухающими компонентами с существенно различающимися скоростями затухания:

$$\psi_1(t) = \psi_1(0)\exp(-\lambda_1 t), \quad \psi_2(t) = \psi_2(0)\exp(-\lambda_2 t).$$

Первая компонента является быстрой, однако при $t > 1$ ее значение близко к нулю и по абсолютному значению практически не меняется. Малость быстрой компоненты вместе с ее производными высших порядков указывает на малость локальной погрешности и возможность использования относительно больших значений шага, тем более, что вторая компонента остается медленной и также допускает использование крупных шагов численного интегрирования.

Тем не менее, условия устойчивости для явных методов приводят к ограничению на размер шага вида

$$\tau \leqslant 2\|A\|^{-1} = 2\lambda_1^{-1}. \tag{1.54}$$

Если условия устойчивости не выполнено, то стационарное состояние первой компоненты становится неустойчивым, как это показано на рис. 1.1.

Описанная выше ситуация является типичной для линейных систем дифференциальных уравнений вида (1.52), когда для собственных значений матрицы A, $\lambda(A) = \lambda_k$, $k = 1, 2, \ldots, n$, имеет место следующее:

$$\mathrm{Re}(\lambda(A)) > 0, \tag{1.55}$$

$$s = \frac{\max Re(\lambda(A))}{\min Re(\lambda(A))} >> 1. \tag{1.56}$$

Системы дифференциальных уравнений вида (1.52) с плохо обусловленной матрицей A, собственные значения которой характеризуются неравенствами (1.55)–(1.56), принято называть **жесткими**, а величину s — **коэффициент жесткости**. Коэффициент жесткости является количественной характеристикой жестких систем. Чем больше коэффициент жесткости, тем более сложным может оказаться численное интегрирование такой задачи.

Понятие жесткости имеет естественное обобщение на случай нелинейных систем

$$\frac{du}{dt} = -f(t, u), \tag{1.57}$$

для которых, вместо матрицы A, неравенства (1.55)–(1.55), следует рассматривать применительно к собственным значениям матрицы Якоби, порождаемой вектор–функцией правой части системы — $f(t, u)$.

Для того, чтобы избежать дополнительных ограничений на размер шага, не связанных с требованиями точности, следует использовать безусловно устойчивые (A-устойчивые) неявные методы. Каждый шаг неявных методов требует больших вычислительных затрат, однако, благодаря возможности использовать более крупный размер шага без потери устойчивости, такие методы в ряде случаев оказываются более предпочтительными по сравнению с явными. В рамках неявной многошаговой схемы Адамса – Моултона мы можем получить абсолютно устойчивый метод не превосходящий второго порядка точности.

Для улучшения устойчивости многошагового метода более перспективным представляется использование чисто неявной схемы, основанной на **формулах дифференцирования назад** (**backward differentiation formulae**). Данный класс многошаговых методов известен также как **методы Гира**:

$$\sum_{m=0}^{M} a_m U_{k-m} = \tau f(t_k, U_k). \tag{1.58}$$

Для получения минимальной локальной ошибки коэффициенты a_m находятся как решение следующей системы алгебраических уравнений:

$$\begin{array}{lllll} a_1 & + & 2a_2 & + \ldots + & Ma_m = -1, \\ a_1 & + & 2^2 a_2 & + \ldots + & M^2 a_m = 0, \\ \ldots & & & & \\ a_1 & + & 2a_M & + \ldots + & M^M a_m = 0, \end{array} \tag{1.59}$$

$$a_0 = -\sum_{m=1}^{M} a_m.$$

Методы Гира имеют глобальную погрешность $O(\tau^M)$ и сохраняют безусловную устойчивость для действительных λ вплоть до $M = 6$. Однако при $M > 6$ этот метод является абсолютно неустойчивым, т.е. устойчивость не может быть обеспечена сколь угодно малым размером шага τ.

Одношаговым аналогом метода Гира является неявный метод Эйлера (1.19). Наиболее значимые с практической точки зрения формулы дифференцирования назад имеют следующий вид и локальную погрешность:

$$\begin{aligned}
&\frac{3}{2}U_k - 2U_{k-1} + \frac{1}{2}U_{k-2} = \tau f_k, && \delta_u = -\frac{2\tau^3}{9}u''',\\
&\frac{11}{6}U_k - 3U_{k-1} + \frac{3}{2}U_{k-2} - \frac{1}{3}U_{k-3} = \tau f_k, && \delta_u = -\frac{3\tau^4}{22}u^{(4)},\\
&\frac{25}{12}U_k - 4U_{k-1} + 3U_{k-2} - \frac{4}{3}U_{k-3} + \frac{1}{4}U_{k-4} = \tau f_k, && \delta_u = -\frac{12\tau^5}{125}u^{(5)},\\
&\frac{137}{60}U_k - 5U_{k-1} + 5U_{k-2} - \frac{10}{3}U_{k-3} + \frac{15}{12}U_{k-4} - \frac{1}{5}U_{k-5} = \tau f_k, && \delta_u = -\frac{10\tau^6}{137}u^{(6)}.
\end{aligned} \tag{1.60}$$

Область устойчивости многошаговых методов Гира является неограниченной, как можно заметить из иллюстрации, представленной на рис. 1.4.

Неявные формулы дифференцирования назад приводят к системам нелинейных алгебраических уравнений вида

$$F(U_k) = 0, \quad F(U_k) = U_k - a_0^{-1}\tau f(t_k, U_k) + a_0^{-1}\sum_{m=1}^{M} a_m U_{k-m}, \tag{1.61}$$

для решения которых может быть использован итерационный метод Ньютона:

$$U_k^{(s+1)} = U_k^{(s)} - \left[I - \tau a_0^{-1}\frac{\partial f_k^s}{\partial U}\right]^{-1}\left[U_k^{(s)} - \tau a_0^{-1} f_k^s + a_0^{-1}\sum_{m=1}^{M} a_m U_{k-m}\right]. \tag{1.62}$$

Здесь $\dfrac{\partial f_k^s}{\partial U}$ — матрица Якоби, соответствующая вектор–функции правой части системы $f_k^s = f(U_k^s, t_k)$. Сходимость метода Ньютона (1.62) может быть обеспечена без существенных ограничений размера шага τ при достаточно близком начальном приближении $U_k^{(0)}$, например, используя какой - либо явный метод в качестве предиктора.

Другой подход к решению жестких систем основан на использовании **неявных методов Рунге – Кутты**. Напомним, что традиционно методы Рунге – Кутты явного типа (1.23) могут быть определены набором коэффициентов, которые удобно представить в виде таблицы Бутчера (1.23). В случае явных методов таблица Бутчера имеет треугольный вид. Естественным обобщением явных формул являются их неявные аналоги с полной матрицей коэффициентов A, $a_{km} \neq 0$, $m \geqslant k$:

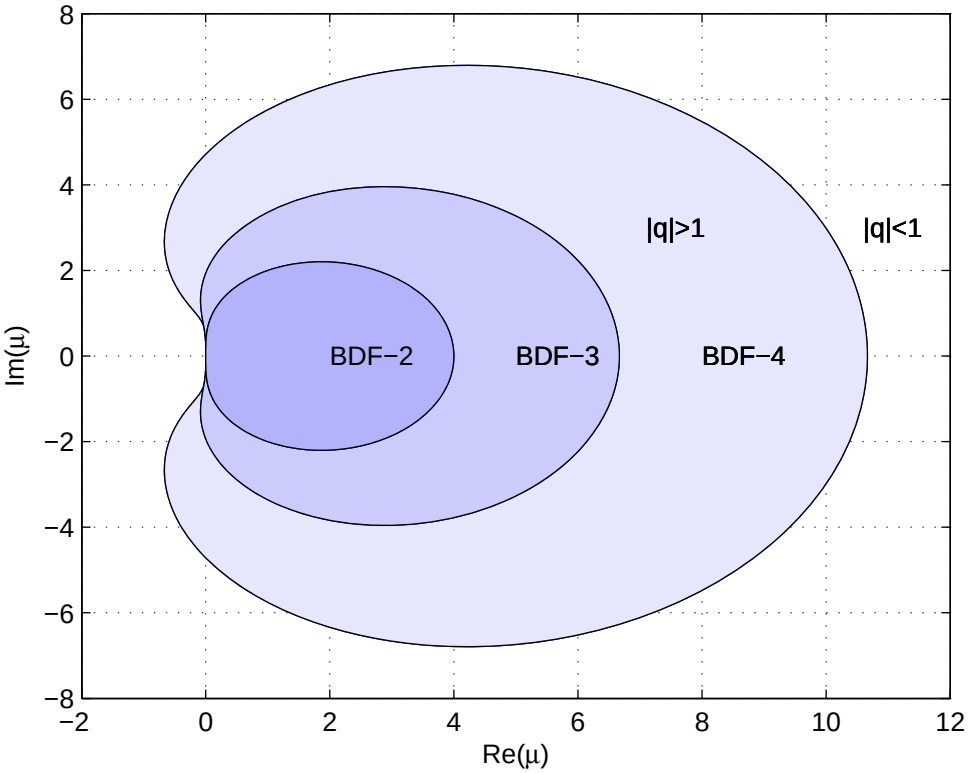

Рис. 1.4. Границы областей устойчивости чисто неявных формул дифференцирования назад для методов Гира второго – четвертого порядка. Отметим, что области устойчивости лежат снаружи закрашенных областей.

$$
\begin{aligned}
K_1 &= \tau f(t_k + c_1\tau,\, U_k + \sum_{i=1}^{s} a_{1i}K_i),\\
K_2 &= \tau f(t_k + c_2\tau,\, U_k + \sum_{i=1}^{s} a_{2i}K_i),\\
&\cdots\cdots\cdots\cdots\\
K_s &= \tau f(t_k + c_s\tau,\, U_k + \sum_{i=1}^{s-1} a_{si}K_i),
\end{aligned}
\tag{1.63}
$$

$$
U_{k+1} = U_k + \sum_{m=1}^{s} b_m K_m, \tag{1.64}
$$

where

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array} \tag{1.65}$$

Значения компонент c, b и A определяются требованием минимизации локальной погрешности метода. Например, двухстадийный неявный метод Рунге – Кутты максимальной точности, основанный на квадратурных формулах Гаусса – Лежандра, определяется следующим набором коэффициентов

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cc} \frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{3-2\sqrt{3}}{12} \\ \frac{3+\sqrt{3}}{6} & \frac{3+2\sqrt{3}}{12} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{1.66}$$

В отличие от явного метода Рунге – Кутты, где вспомогательные значения K_m, вычисляются последовательно, используя уже вычисленные значения $K_1, \dots K_{m-1}$, для определения последующих, $m = 1, 2, \dots s$, в неявном методе для вычисления значений K_m необходимо решить соответствующую систему алгебраических уравнений (1.63), вообще говоря, нелинейную. Размерность систем алгебраических уравнений, требующих решения на каждом шаге, пропорционально количеству уравнений дифференциальной задачи, и количеству стадий неявного метода. Таким образом, неявные методы Рунге – Кутты с вычислительной точки зрения более затратны. Ситуация несколько упрощается в случае неявного диагонального метода Рунге – Кутты, когда матрица коэффициентов имеет нижний треугольный вид с ненулевыми диагональными значениями: $a_{km} \neq 0$, если $m \leqslant k$, и $a_{km} = 0$ в противном случае. При таких условиях каждое значение K_m находится независимо от остальных.

Среди достоинств неявных методов Рунге – Кутты, прежде всего следует отметить его преимущества в устойчивости, что делает весьма привлекательным использование этого класса методов для численного анализа жестких систем. В случае тестовой задачи (1.41) метод Рунге – Кутты (1.63) – (1.65) приводит к рекурсивному равенству

$$U_k = R(\tau\lambda)U_n, \tag{1.67}$$

где $R(z)$ имеет вид

$$R(z) = 1 + zb^T(I - zA)^{-1}e. \tag{1.68}$$

Здесь $e = (1, 1, \dots 1)^T$, I — единичная матрица. Условие устойчивости определяется неравенством:

$$|R(z)| \leqslant 1. \tag{1.69}$$

Как будет показано ниже, в рамках неявного метода Рунге – Кутты возможно построение A-устойчивых методов выше второго порядка точности.

Следует также отметить, что порядок точности неявного метода Рунге – Кутты при одинаковом числе стадий превосходит точность явной схемы. Так, например, двухстадийный неявный метод (1.69) имеет четвертый порядок точности, в то время как явная схема достигает только второго порядка. В общем случае, s-стадийный неявный метод Рунге –Кутты имеет порядок точности вплоть до $O(\tau^{2s})$ в то время как для явных аналогов порядок точности не превосходит $O(\tau^s)$. Для диагональной неявной схемы Рунге – Кутты максимальный порядок точности не превышает $O(\tau^{s+1})$. Пример такой двухстадийной диагональной неявной схемы имеет следующий набор коэффициентов:

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} = \begin{array}{c|cc} \gamma & \gamma & 0 \\ 1-\gamma & 1-2\gamma & \gamma \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{1.70}$$

with $\gamma = \dfrac{3+\sqrt{3}}{6}$

Пример 1.3. Сравним область устойчивости явного и неявного метода Рунге – Кутты четвертого порядка точности, представленных коэффициентами (1.26) и (1.66) соответственно. Пример расчета области устойчивости методов Рунге – Кутты на основе условия (1.69) вместе с программной реализацией и результатами численных расчетов представлены ниже.

```
%% область устойчивости явного и неявного методов РК
%% A,b,e коэффициенты явной четырехстадийной схемы РК %%%%%
%% A1,b1,e1 коэффициенты неявной двухстадийной схемы РК  %%%
x = -4:0.01:1; y = -3.5:0.01:3.5; [X,Y] = ndgrid(x,y);
A = [0 0 0 0; 1/2 0 0 0;0 1/2 0 0;0 0 1 0];
A1 = [1/4 (3-2*sqrt(3))/12; (3+2*sqrt(3))/12 1/4];
b = [1 2 2 1]/6; b1 = [1/2 1/2];
e = [1 1 1 1]; e1 = [1 1];
E = eye(4); E1 = eye(2);
R = zeros(size(X)); R1 = R;
for k = 1:length(x);
for m = 1:length(y);
z = X(k,m)+1i*Y(k,m);
R(k,m) = abs(1+z*b*(inv(E-z*A)*e'));
R1(k,m) = abs(1+100*z*b1*(inv(E1-100*z*A1)*e1'));
end
```

```
end
%% граница области устойчивости задана
%% уравнением: 1-|RE(z)|=0 и отображается изолинией
subplot(2,1,1), [C, H]= contourf(X,Y,1-R,[0 3])
colormap([0.9 0.9 0.99])
title('The Stability Domain of the 4-th order Explicit RK')
text(-0.7,0.9,'|R(z)|<0')
text(0.3,0.9,'|R(z)|>0')
xlabel('Re(z)')
ylabel('Im(z)')
grid on
subplot(2,1,2), [C, H]= contourf(100*X,100*Y,1-R1,[0 3]);
colormap([0.9 0.9 0.99]);
text(-90,90,'|R(z)|<0');
text(30,90,'|R(z)|>0')
xlabel('Re(z)')
ylabel('Im(z)')
title('The Stability Domain of the 4-th order Implicit RK')
grid on
```

Рассмотренный пример показывает, что неявный метод Рунге – Кутты четвертого порядка имеет неограниченную область устойчивости, покрывающую всю левую полуплоскость комплексной плоскости, т.е. данный метод является A-устойчивым.

Несмотря на превосходные свойства устойчивости и высокую точность неявные методы Рунге – Кутты имеют серьезный недостаток, связанный с высокой вычислительной сложностью реализации. По этой причине практический интерес привлекают преимущественно варианты неявной схемы невысокого порядка точности для решения жестких систем при умеренных требованиях к точности результатов. В частности, некоторые варианты неявных схем реализованы в функциях MATLAB[1] **ode23tb**, **ode23t** и **ode23s**. Многошаговые методы переменного порядка, включая чисто неявные формулы дифференцирования назад, реализованы в функции **ode15s**. Сравнительный анализ вычислительной эффективности упомянутых методов рассмотрен на примере решения тестовой задачи ниже.

Пример 1.4. В качестве тестовой задачи для сравнительной оценки эффективности функций MATLAB, реализующих различные методы решения задачи Коши рассмотрим систему Ван дер Поля:

$$\begin{aligned} \frac{du}{dt} &= v, \\ \frac{dv}{dt} &= \mu(1-u^2)v - u, \end{aligned} \tag{1.71}$$

$$u(0) = -2, \quad v(0) = 0, \quad t = [0, T_\mu]. \tag{1.72}$$

Использованы значения параметров $\mu = 1;\ 10;\ 100;\ 1000$ и $T_\mu = 5\mu$. Рассмотренная задача является классическим примером жесткой системы с коэффициентом

[1] См. подробнее Ashino, R., Nagase, M., and Vaillancourt, R. (2000). Behind and beyond the MATLAB ODE suite. Computers and Mathematics with Applications, 40(4), 491-512.

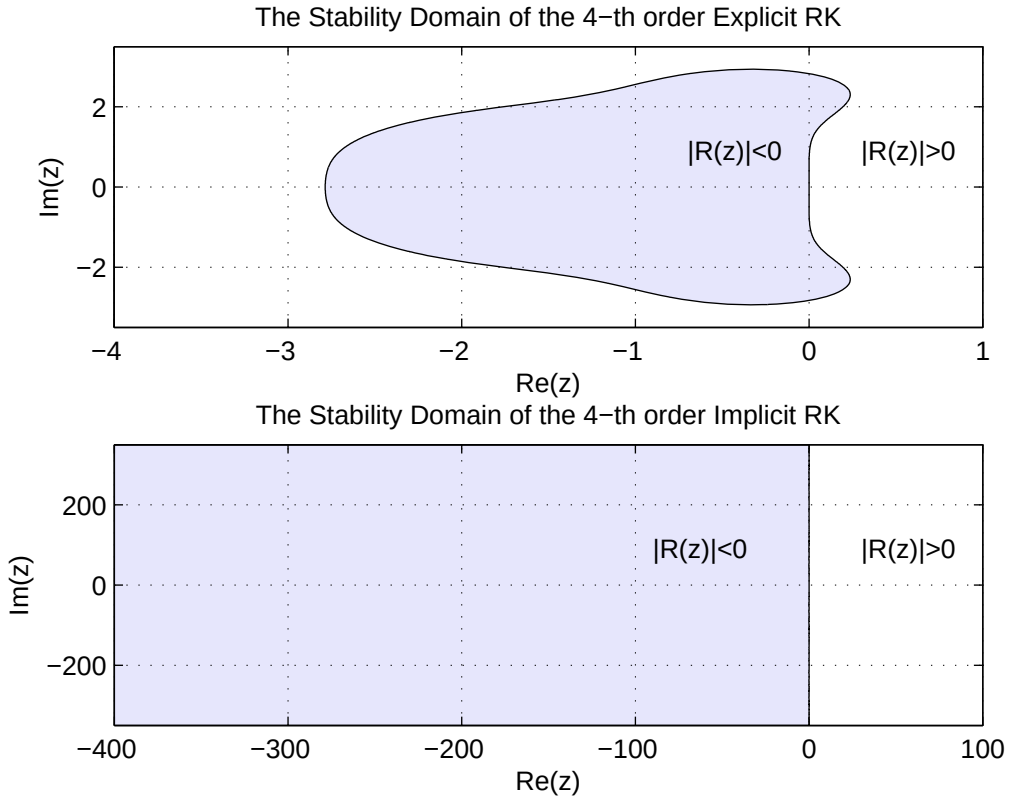

Рис. 1.5. Области устойчивости явного и неявного методов Рунге – Кутты четвертого порядка точности.

жесткости, определяемым значением параметра μ. В области значений параметра от $\mu = 1$ до $\mu = 1000$ мы можем проследить поведение различных методов при изменении свойств системы от не жесткой до сильно жесткой.

Наша цель состоит в сравнении эффективности различных неявных методов применительно к анализу жестких систем ОДУ. Полезно также проиллюстрировать некоторые возможности управления параметрами численных методов с целью достижения их максимальной эффективности.

Программная реализация рассмотренной задачи и результаты численных экспериментов представлены ниже.

```
%% Жесткая система Ван дер Поля system.%%%%%%%%%%%%%%%%%%%%%%%%
%% неявный метод РК 2-3-го порядка  (функция ODE23)
%% Формулы обратного дифференцирования 1-5го порядка (функция ODE15s )
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear
settings = 'default';
settings = 'precalc';
k = 0;  Mu = [1 10 100 1000];
```

```
for M = Mu
k = k+1;  mu = M;
T=10*mu
options = odeset('RelTol',1.e-7); % Relative Tolerance
%%   Функция правой части и матрица Якоби
dydt = @(t,y)[y(2); mu*(1-y(1)^2)*y(2)-y(1)];
dfdt = @(t,y)[0 1; -2*mu*y(1)*y(2)-1 mu*(1-y(1)^2)];
if settings == 'default'
else
options = odeset('Jacobian',dfdt ); % Jacobian func.
end
```

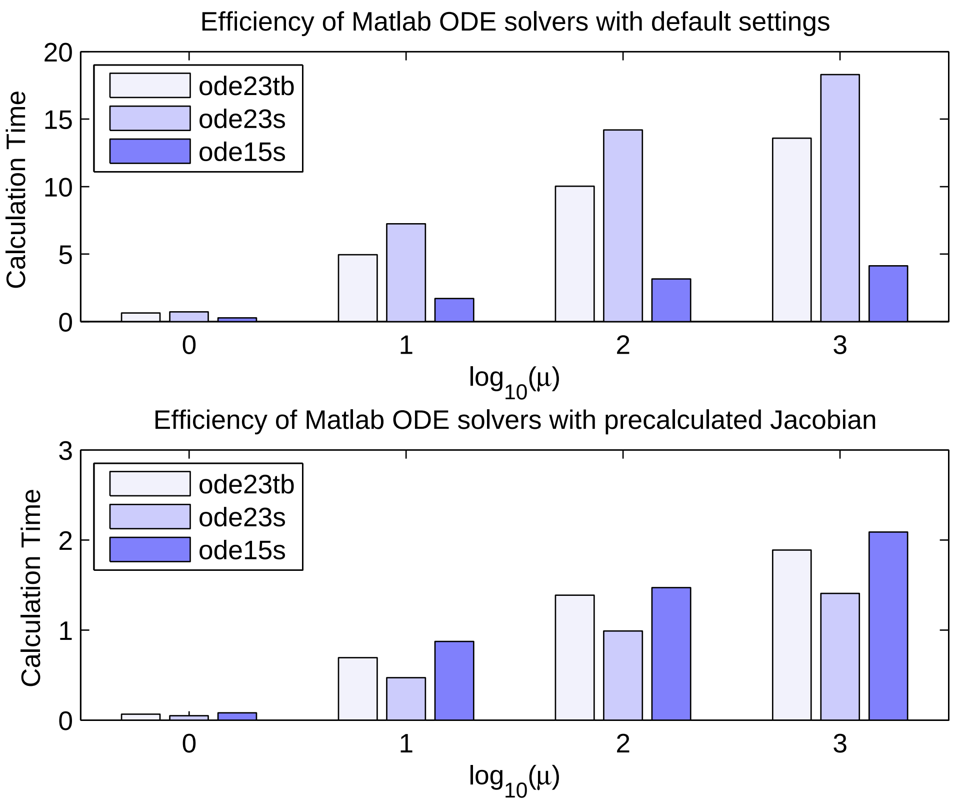

Рис. 1.6. Сравнение вычислительной эффективности функций MATLAB на примере моделирования жесткой системы Ван дер Поля (1.71), (1.71) при использовании настроек численного метода по умолчанию и предварительно заданным явным видом матрицы Якоби для вектор-функции правой части системы

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tic
```

```
sol_bdf = ode15s(dydt,[0 T],[-2 0],options); % BDF
time_BDF(k) = toc
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tic
sol_irk = ode23s(dydt,[0 T],[-2 0],options); % IRK
time_IRK(k) = toc
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tic
sol_irk = ode23tb(dydt,[0 T],[-2 0],options); % IRK
time_IRB(k) = toc
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
end
if settings == 'default'
subplot(2,1,1),
bar(log10(Mu),[time_IRB(:) time_IRK(:), time_BDF(:)])
title('Efficiency of Matlab ODE solvers with default settings')
else
subplot(2,1,2),
bar(log10(Mu),[time_IRB(:) time_IRK(:), time_BDF(:)])
title('Efficiency of Matlab ODE solvers with precalculated Jacobian')
end
colormap([0.95 0.95 0.99; 0.8 0.8 0.99;0.5 0.5 0.99]);
legend('ode23tb','ode23s','ode15s','Location', 'NorthWest')
xlabel('log_{10}(\mu)');  ylabel('Calculation Time ')
```

Результаты численного моделирования, представленные на рис. 1.4, позволяют заметить, что в случае, когда Якобиан системы не задан аналитически, существенное преимущество демонстрирует функция **ode15s**, реализующая неявные многошаговые методы переменного порядка. Эффективность многошаговых методов практически не изменяется при использовании аналитического вида Якобиана для реализации неявной нелинейной схемы, в то время как неявные методы Рунге – Кутты, реализованные в функциях **ode23s** и **ode23tb** получают при этом почти десятикратное ускорение.

Упражнение 1.5.

1. Сравните области устойчивости неявного и неявно-диагонального метода Рунге – Кутты, заданных коэффициентами (1.66) и (1.70).

2. Сравните области устойчивости явных методов Рунге – Кутты четвертого и пятого порядков точности.

3. Вычислите коэффициенты неявного метода Гира 7-го порядка. Определите область устойчивости данного метода.

4. Внося необходимые модификации в программу рассмотренного выше примера сравните эффективность функций MATLAB **ode23tb**, **ode23s** , **ode15s**, **ode45** используя в качестве тестовой задачи систему Лоренца (??), (??). Сравните полученные результаты оценки эффективности на примерах систем Ван дер Поля и Лоренца.

Глава 2

Краевые задачи

2.1. Постановка задачи

При математическом моделировании инженерно-физических задач возникает необходимость определить дополнительные условия для решения систем дифференциальных уравнений при различных значениях независимой переменной. Для простоты рассмотрим пример дифференциального уравнения второго порядка

$$\frac{d^2u}{dx^2} + a(x)u = f(x), \quad x \in (0,1), \tag{2.1}$$

с нулевыми краевыми условиями на концах отрезка:

$$u(0) = 0, \quad u(1) = 0. \tag{2.2}$$

Задачи такого типа, когда дополнительные условия задаются в разных точках, называются **краевыми задачами**.

Задача (2.1)–(2.2) допускает эквивалентное представление в виде системы двух дифференциальных уравнений первого порядка:

$$\begin{cases} \dfrac{dv}{dx} + a(x)u = f(x), \\ \dfrac{du}{dx} = v, \end{cases} \tag{2.3}$$

Формулировка краевой задачи в виде системы дифференциальных уравнений первого порядка в некоторых случаях представляется предпочтительной, особенно если искомые функции такой системы наделены конкретным смыслом. В общем случае двухточечная краевая задача для системы дифференциальных уравнений N-го порядка имеет следующий вид

$$\frac{d\mathbf{u}}{dx} = \mathbf{f}(\mathbf{u}, x), \quad \mathbf{f}(\mathbf{u}, x), = [f_1(\mathbf{u}, x), f_2(\mathbf{u}, x), \ldots, f_N(\mathbf{u}, x)]^T, \quad \mathbf{u} = [u_1, u_2, \ldots, u_N]^N \tag{2.4}$$

с краевыми условиями

$$\mathbf{B}^L(\mathbf{u})\,|_{x=L} = 0, \quad \mathbf{B}^R(\mathbf{u})\,|_{x=R} = 0\ , \tag{2.5}$$

где $x = L$ и $x = R$ — левая и правая границы отрезка, $\mathbf{B}^L$, $\mathbf{B}^R$ — вектор-функции,

$$\mathbf{B}^L = [b_1^L(\mathbf{u}), b_2^L(\mathbf{u}), \ldots, b_n^L(\mathbf{u})], \quad \mathbf{B}^R = [b_1^R(\mathbf{u}), b_2^R(\mathbf{u}), \ldots, b_{N-n}^R(\mathbf{u})],$$

такие, что общее число условий (2.5) равняется числу уравнений системы (2.4).

2.2. Метод конечных разностей и конечных элементов для решения краевых задач

Основная идея большинства методов численного анализа дифференциальных уравнений состоит в переходе от рассмотрения непрерывной задачи в бесконечномерном функциональном пространстве V к дискретной задаче, формулируемой в конечномерном векторном пространстве V_N. Среди наиболее универсальных подходов к решению краевых задач для дифференциальных уравнений следует в первую очередь отметить **методы конечных разностей** и **конечных элементов**.

В методе конечных разностей, в отличие от исходной дифференциальной задачи, искомое решение является не функцией, заданной в некоторой ограниченной области, а множество значений этой функции, определенных в конечном числе точек, принадлежащих области определения решения.

Теоретическую основу метода конечных элементов составляют слабая формулировка задачи, метод Галеркина и представление решения в виде линейной комбинаций финитных базисных функций $\Psi_n(x)$, $n = 1, \ldots, N$, имеющих компактный носитель и удовлетворяющих краевым условиям:

$$U(x) = \sum_{n=1}^{N} c_n \Psi_n(x), \quad \Psi_n(x) \neq 0,\ |x - x_n| \leqslant \Delta x, \quad \Psi_n(x) = 0,\ |x - x_n| > \Delta x, \tag{2.6}$$

Несмотря на принципиальное отличие методов конечных разностей и конечных элементов, построенные в рамках данных подходов дискретные модели могут при определенных условиях совпадать. Рассмотрим подробнее эти два метода на примере решения дифференциального уравнения второго порядка на единичном отрезке с нулевыми краевыми условиями Дирихле:

$$\frac{d}{dx}\lambda(x)\frac{du}{dx} - q(x)u = f(x), \quad x \in (0,1), \tag{2.7}$$

$$u(0) = 0, \quad u(1) = 0. \tag{2.8}$$

Для построение дискретной разностной модели в области $\Omega = (0,1)$ определим множество точек $\omega_h = \{x_0, x_1, \ldots, x_N\}$, разбивающее область определения решения дифференциальной задачи на некоторое число интервалов. Как правило $0 \leqslant x_0 < x_1 < \ldots < x_{N-1} < x_N \leqslant 1$. Множество ω_h называется **сеткой**, а сами

точки x_k, $k = 0, \ldots, N$ — **узлы сетки** . Расстояние между соседними узлами называют **шагом сетки**, $h_k = x_k - x_{k-1}$. Для простоты мы ограничимся рассмотрением случая **равномерной сетки**, когда все шаги одинаковы: $h_1 = h_2 = \ldots h_n = h$. Для точного и приближенного решений в узлах сетки мы будем использовать соответственно следующие обозначения: $u(x_k) = u_k$, $U(x_k) = U_k$.

Конечно–разностный метод основан на замене производных дифференциальной задачи на соответствующие разностные приближения. Мы будем использовать следующие приближенные соотношения для дискретного представления производных на сетке:

$$\frac{du(x_k)}{dx} = \frac{u_{k+1} - u_{k-1}}{2h} + \frac{h^2}{6}u''' + O(h^4), \tag{2.9}$$

$$\frac{d^2u(x_k)}{dx^2} = \frac{u_{k+1} - 2u_k + u_{k-1}}{h^2} + \frac{h^2}{12}u^{(4)} + O(h^4), \tag{2.10}$$

$$\frac{d}{dx}\lambda(x_k)\frac{du(x_k)}{dx} = \frac{\lambda_{k+1/2}\left(u_{k+1} - u_k\right) - \lambda_{k-1/2}\left(u_k - u_{k-1}\right)}{h^2} + O(h^2), \tag{2.11}$$

где $\lambda_{k\pm1/2} = \lambda((x_k + x_{k\pm1})/2)$ или $\lambda_{k\pm1/2} = \left(\lambda(x_k) + \lambda(x_{k\pm1})\right)/2$.

Уравнения (2.9)–(2.11) получены на основе представления решения и коэффициентов дифференциальной задачи отрезком степенного ряда:

$$u_{k\pm1} = u_k \pm \frac{u'(x_k)}{1!}h + \frac{u''(x_k)}{2!}h^2 \pm \frac{u'''(x_k)}{3!}h^3 + \frac{u^{(4)}(x_k)}{4!}h^4 + O(h^5). \tag{2.12}$$

$$\lambda_{k\pm1/2} = \lambda_k \pm \frac{\lambda'(x_k)}{2 \cdot 1!}h + \frac{\lambda''(x_k)}{4 \cdot 2!}h^2 \pm \frac{\lambda'''(x_k)}{8 \cdot 3!}h^3 + \frac{\lambda^{(4)}(x_k)}{16 \cdot 4!}h^4 + O(h^5). \tag{2.13}$$

Подстановка выражений (2.11) в (2.7) приводит к следующим уравнениям

$$\frac{\lambda_{k+1/2}\left(u_{k+1} - u_k\right) - \lambda_{k-1/2}\left(u_k - u_{k-1}\right)}{h^2} - q(x_k)u_k - f(x_k) = R(h), \quad x_k \in \omega_h, \tag{2.14}$$

где $R(h) = O(h^2)$ — **погрешность аппроксимации**, возникающая в силу приближенности разностного представления производных. Отметим, что если в разложении (2.12)–(2.13) производные высших порядком существуют и ограничены, то погрешность аппроксимации стремиться к нулю при убывании шага сетки:

$$\lim_{h \to 0} R(h) = 0.$$

Если шаг сетки достаточно мал мы можем пренебречь погрешностью аппроксимации, что приводит к следующему уравнению для приближенного решения задачи (2.7))–(2.8):

$$\frac{\lambda_{k+1/2}\left(U_{k+1} - U_k\right) - \lambda_{k-1/2}\left(U_k - U_{k-1}\right)}{h^2} - q(x_k)U_k - f(x_k) = 0, \quad k = \overline{1, N-1}, \tag{2.15}$$

$$U_0 = U_N = 0. \tag{2.16}$$

Уравнения вида (2.15)–(2.16), полученные заменой производных дифференциальной задачи с помощью соответствующих конечных разностей называются **разностной схемой**.

Разностные схемы являются приближенными дискретными моделями дифференциальных задач. В рассмотренном случае данная модель имеет вид системы линейных алгебраических уравнений с трехдиагональной матрицей. Отметим, что различие между конечно–разностным уравнением (2.14) для точного решения исходной дифференциальной задачи и разностной схемой (2.15)–(2.16) определяется величиной $R(h)$ которая может быть интерпретирована как **невязка** — дисбаланс в приближенном разностном уравнении при подстановке в него точного решения исходной дифференциальной задачи. Мы будем говорить, что разностная схема **согласована** или разностная схема **аппроксимирует** дифференциальную задачу, если невязка (погрешность аппроксимации) разностной схемы на решении соответствующей дифференциальной задачи стремится к нулю при шаге сетки стремящемся к нулю.

Как правило, порядок малости невязки разностной схемы характеризует скорость сходимости ее решения. Напомним, что **скорость сходимости** или **порядок точности** численного метода — это асимптотическая скорость убывания разницы между приближенным решением U_k и его точными значениями $u_k = u(x_k)$ в узлах сетки при шаге сетки стремящемся к нулю. Скорость сходимости является одной из важнейших характеристик разностных схем — количественной мерой их точности. Говоря о том, что численный метод имеет n-й порядок точности мы подразумеваем, что при достаточно малом h погрешность приближенного решения $\delta(x_k) = U(x_k) - u(x_k)$, $x_k \in \omega_h$, имеет такой же порядок малости: $\|\delta\| \leqslant Ch^n = O(h^n)$, где C — постоянная не зависящая от h.

Основное отличие метода конечных элементов от разностных методов состоит в так называемой слабой формулировке дифференциальной задачи. Несложно заметить в уравнениях (2.9)–(2.11), что для согласованности (аппроксимации) разностной схемы требуется существование и ограниченность производных высшего порядка. Например, для получения второго порядка аппроксимации разностной схемы (2.14) мы должны иметь как минимум четырежды дифференцируемое решение дифференциальной задачи, в то время как соответствующее дифференциальное уравнение содержит только вторые производные. Такого рода "сильные"требования, предъявляемые к дифференцируемости решения задачи, могут быть преодолены при использовании слабой формулировки. Слабая формулировка задачи получается путем умножения дифференциального уравнения (2.17) на некоторую дифференцируемую тестовую функцию $\Psi(x)$ с последующим интегрированием полученного равенства, используя формулы интегрирования по частям:

$$\int_0^1 \lambda(x)\frac{du}{dx}\frac{d\Psi(x)}{dx}\,dx + \int_0^1 q(x)u(x)\Psi(x)\,dx + \int_0^1 f(x)\Psi(x)\,dx = 0. \qquad (2.17)$$

Интегральное уравнение (2.17) называется слабой постановкой задачи, а функция $u(x)$, удовлетворяющее данному уравнению, называется слабым решением дифференциальной задачи (2.7)–(2.8). При построении дискретной модели на основе методов конечных элементов мы также используем дискретизацию области определения решения, разбивая ее на подобласти множеством узлов сетки. Кроме того, мы

определяем множество базисных функций — **функций формы**. Обычно в качестве базисных функций используются кусочно–полиномиальные функции (B-сплайны) отличные от нуля в пределах ячейки сетки. Приближенное решение задачи находится в виде линейной комбинации базисных функций (2.6), с коэффициентами c_n, которые определяются на основе **метода Галеркина**: коэффициенты c_n таковы, что невязка дискретной модели ортогональна всем базисным функциям. Фактически метод Галеркина соответствует ортогональной проекции решения задачи на линейную оболочку базисных функцийΨ_n, $n = 1, \ldots, N$, что обеспечивает минимальную погрешность приближения в Гильбертовом пространстве.

Ортогональность понимается здесь в смысле равенства нулю скалярного произведения векторов. Скалярное произведение функций на отрезке $[a, b]$ определяется следующим образом:

$$(\Psi_n, \Psi_m) = \int_a^b \Psi_n(x)\Psi_m(x)\,dx. \tag{2.18}$$

Две функции считаются ортогональными если их скалярное произведение равно нулю:

Для вычисления коэффициентов c_n, $n = 1, \ldots, N$, подставляем решение (2.6) в уравнение (2.7):

$$\sum_{n=1}^{N}\left\{\frac{d}{dx}\lambda(x)\frac{d\Psi_n(x)}{dx}c_n - q(x)c_n\Psi_n(x) - f(x)\right\} = R(x). \tag{2.19}$$

После этого умножим полученное равенство (2.19) поочередно на базисные функции $\Psi_m(x)$, $m = 1, \ldots, N$ и проинтегрируем полученные равенства:

$$\sum_{n=1}^{N} c_n\left\{\int_0^1 \frac{d}{dx}\lambda(x)\frac{d\Psi_n(x)}{dx}\Psi_m(x)\,dx - \int_0^1 q(x)\Psi_n(x)\Psi_m(x)dx\right\} - \int_0^1 f(x)\Psi_m(x)\,dx - \int_0^1 R(x)\Psi_m(x)\,dx = 0, \quad m = 1, \ldots, N. \tag{2.20}$$

Последний интеграл в уравнении (2.20) равен нулю в силу требования ортогональности невязки и пробных функций. Вычисление первого интеграла в уравнении (2.20) по частям приводит к выражению

$$\sum_{n=1}^{N}\left\{\int_0^1 \lambda(x)\frac{d\Psi_m(x)}{dx}\frac{d\Psi_n(x)}{dx}\,dx + \int_0^1 q(x)\Psi_n(x)\Psi_m(x)\,dx\right\}c_n = -\int_0^1 f(x)\Psi_m(x)\,dx. \tag{2.21}$$

Уравнения (2.21), $m = 1, \ldots, N$, соответствуют слабой постановке задачи (2.7)–(2.8). Коэффициенты c_n находятся как решение следующей системы линейных алгебраических уравнений:

$$Ac = g; \tag{2.22}$$

where $c = [c_1, c_2, \ldots, c_N]^T$, $g = [g_1, g_2, \ldots, g_N]^T$,

$$\{A_{mn}\} = \int_0^1 \lambda(x) \frac{d\Psi_m(x)}{dx} \frac{d\Psi_n(x)}{dx}\, dx + \int_0^1 q(x)\Psi_n(x)\Psi_m(x)\, dx, \tag{2.23}$$

$$g_m = \int_0^1 f(x)\Psi_m(x)\, dx. \tag{2.24}$$

В общем случае матрица системы уравнений (2.22) является симметричной и полной (отсутствует преобладание нулевых элементов). Однако, если используемые базисные функции $\Psi_m(x)$ имеют компактный носитель на отрезке $x \in [x_{m-1}, x_{m+1}]$, то матрица системы (2.22) будет иметь трехдиагональную структуру, аналогичную структуре матрицы разностной модели: $\{A_{m,n}\} = 0$, когда $|n - m| \geqslant 1$.

Пример 2.1. Сравним возможности методов конечных разностей и конечных элементов на примере решения задачи (2.7)–(2.8) для следующих входных данных: $\lambda(x) = 1$, $q(x) = 0$, $f(x) = 32(x - 0.5)^4 - 4(x - 0.5)^2$.Разностный метод в данном случае приводит к следующей системе линейных алгебраических уравнений

$$AU = b, \tag{2.25}$$

где $b = [f(x_1), f(x_2), \ldots, f(x_{N-1})]^T, U = [U_1, U_2, \ldots, U_{N-1}]^T$ — приближенное решение задачи в точках равномерной сетки, A — симметричная трехдиагональная матрица размерности $(N-1) \times (N-1)$:

$$A = \frac{1}{h^2} \begin{bmatrix} -2 & 1 & & & & & \\ 1 & -2 & 1 & & & 0 & \\ & \ddots & \ddots & \ddots & & & \\ & & & & & & \\ & & & & & & \\ & 0 & & & \ddots & \ddots & \ddots \\ & & & & & 1 & -2 & 1 \\ & & & & & & 1 & -2 \end{bmatrix} \tag{2.26}$$

Отметим, что в силу нулевых краевых условий мы формулируем разностную задачу только для внутренних точек сетки (x_k, $k = 1, \ldots, N-1$). Для учета ненулевых условий Дирихле ($u(0) = \alpha$, $u(1) = \beta$) следует воспользоваться соответствующей коррекцией первой и последней компонент вектора правой части системы:: $b(1) = f(x_1) - \alpha/h^2$, $b(N-1) = f(x_{N-1}) - \beta/h^2$.

В случае краевых условий Неймана, входящие в них производные следует аппроксимировать конечными разностями и приближенное решение в граничных точках вместе с соответствующими уравнениями включаются в дискретную разностную модель. Например, в случае **краевых условий Неймана** общего вида

$$\left[\lambda(x)\frac{\partial u}{\partial x} + \alpha_0(x)u + \gamma_0(x)\right]_{x=0} = 0, \quad \left[\lambda(x)\frac{\partial u}{\partial x} + \alpha_L(x)u + \gamma_L(x)\right]_{x=L} = 0, \tag{2.27}$$

наиболее подходящим представляется использование модифицированной сетки $\overline{\omega}_h$, в которой первый и последний узлы располагаются на расстоянии в половину шага снаружи области определения решения $x \in (0, L)$:

$$\overline{\omega}_h = \{x_k = (k - 1/2)h, \quad k = 0, 1, \ldots, N, \quad h = L/(N-1)\}. \tag{2.28}$$

Использование такой сетки с фиктивными узлами позволяет легко получить второй порядок аппроксимации краевых условий, поскольку граничные точки в данном случае является средними точками крайних интервалов сетки, в которой разностные производные и средние арифметические значения сеточных функций обеспечивают второй порядок точности:

$$\lambda(x_0 + h/2)\,(U_1 - U_0+)\,/h + \alpha_0(x_0 + h/2)\,(U_1 + U_0)\,/2 + \gamma_0(x_0 + h/2). \tag{2.29}$$

$$\lambda(x_N - h/2)\,(U_N - U_{N-1})\,/h + \alpha_0(x_N - h/2)\,(U_N + U_{N-1})\,/2 + \gamma_0(x_N - h/2). \tag{2.30}$$

Эти дополнительные два уравнения должны быть добавлены в качестве первого и последнего уравнений системы (2.25). Как следствие, размерность системы разностных уравнений возрастает с $(N-1) \times (N-1)$ до $(N+1) \times (N+1)$.

Рассмотрим также метод конечных элементов на основе линейных функций формы (линейных B-сплайнов):

$$\Psi_k(x) = \begin{cases} 1 - \dfrac{|x - x_k|}{h}, & x_{k-1} \leqslant x \leqslant x_{k+1}, \\ 0, & x_{k-1} > x > x_{k+1}. \end{cases} \tag{2.31}$$

Для производных данных функций мы имеем

$$\frac{d\Psi_k(x)}{dx} = \begin{cases} -\dfrac{1}{h}, & x_{k-1} \leqslant x \leqslant x_k, \\ \dfrac{1}{h}, & x_k \leqslant x \leqslant x_{k+1}, \\ 0, & x_{k-1} > x > x_{k+1}. \end{cases} \tag{2.32}$$

Для расчета элементов матрицы и компонент вектора правой части системы (2.22) мы можем использовать точные выражения (2.23)–(2.24), или их приближения, например, второго порядка точности:

$$\int_0^1 \lambda(x) \frac{d\Psi_m(x)}{dx} \frac{d\Psi_n(x)}{dx}\, dx \simeq \frac{\lambda(x_m) + \lambda(x_n)}{2} \int_0^1 \frac{d\Psi_m(x)}{dx} \frac{d\Psi_n(x)}{dx}\, dx, \tag{2.33}$$

$$\int_0^1 q(x) \Psi_m(x) \Psi_n(x)\, dx \simeq \frac{q(x_m) + q(x_n)}{2} \int_0^1 \Psi_m(x) \Psi_n(x)\, dx, \tag{2.34}$$

$$\int_0^1 f(x)\Psi_m(x)\,dx \simeq f(x_m)\int_0^1 \Psi_m(x)\,dx. \tag{2.35}$$

Интегралы в формулах (2.33)–(2.35) для конкретного вида функций формы (2.31) могут быть вычислены аналитически:

$$\int_0^1 \frac{d\Psi_m(x)}{dx}\frac{d\Psi_n(x)}{dx} = \begin{cases} -\dfrac{1}{h}, & n = m \pm 1, \\ \dfrac{2}{h}, & n = m, \\ 0, & |n-m| > 1, \end{cases} \tag{2.36}$$

$$\int_0^1 \Psi_m(x)\Psi_n(x)\,dx = \begin{cases} \dfrac{h}{6}, & n = m \pm 1, \\ \dfrac{2h}{3}, & n = m, \\ 0, & |n-m| > 1, \end{cases}, \tag{2.37}$$

$$\int_0^1 \Psi_m(x)\,dx = h. \tag{2.38}$$

Принимая во внимание выражения (2.36)–(2.38), метод конечных элементов приводит к дискретной модели (2.22) для расчета коэффициентов c_k, $k = 2, \ldots, N-1$, и данная модель формально является идентичной разностной модели (2.26). Как следствие, мы имеем $c_k = U_k$. Тем не менее, следует подчеркнуть еще раз, что, несмотря на формальное совпадение значений c_k, и U_k, их смысл остается различным. В методе конечных разностей решение имеет смысл приближенных значений искомого решения в узлах сетки, в то время как в методе конечных элементов решение представлено коэффициентами разложения по базису кусочно линейных функций. Кроме того, метод конечных элементов имеет более широкие возможности повысить точность приближенного решения. Например, если мы вычислим значение интеграла (2.35) аналитически (точно), то метод конечных элементов позволяет получить в рассмотренном случае практически точное решение дифференциальной задачи.

Программная реализация рассмотренного примера и результаты численных экспериментов представлены ниже (см. рис. 2.1).

```
%% Методы конечных разностей и конечных элементов
%% для решения краевой задачи
%% u''=32(x-0.5)^4-4(x-0.5)^2
%% u(0)=u(1)=1.
%% Метод конечных элементов с линейными B-сплайнами
syms t v(t)
```

```
N=31; h = 1/(N-1);
x = 0:h:1;      %
x0 = x(2:end-1);  % Сетка
Nx = length(x0);
%Точное решение на сетке
v = dsolve(diff(v,2)==32*(t-0.5).^4-4*(t-0.5).^2,v(0)==0,v(1)==0);
u=(subs(v,t,x0));
%Вычисление функции правой части
f = @(x)32*(x-0.5).^4-4*(x-0.5).^2;
for m=1:Nx
b0(m) = int((32*(t-0.5)^4-4*(t-0.5)^2)*(1+(t-x0(m))/h),x(m),x0(m))+...
int((32*(t-0.5)^4-4*(t-0.5)^2)*(1-(t-x0(m))/h),x0(m),x(m+2));
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
bb=eval(b0);
e=ones(Nx,1);
A = spdiags([e,-2*e, e],-1:1,Nx,Nx)/h;
C=A\bb(:);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% FDM
A = spdiags([e,-2*e, e],-1:1,Nx,Nx)/h^2;
b= f(x0);
U=A\b(:);
semilogy(x0,abs(C-u(:)),'.-',x0,abs(U-u(:)),'o')
xlabel('x')
ylabel('Err')
legend('FEM','FDM')
FDM_ERR=norm(U-u(:))/norm(u)
FEM_ERR=norm(C-u(:))/norm(u)
```

Применение разностных методов и метода конечных элементов при решении линейной краевой задачи для уравнения второго порядка приводит к дискретным моделям в виде системы линейных алгебраических уравнений с разреженной ленточной матрицей трехдиагонального вида. Такого рода системы могут быть эффективно реализованы с помощью метода Гаусса или его модификаций — метода прогонки. **Вычислительная сложность** реализации такой системы имеет порядок $O(N)$ вместо $O(N^3)$, имеющего место в случае численного анализа линейных систем с полной матрицей.

При решении нелинейных дифференциальных задач соответствующая дискретная модель также будет нелинейной. В этом случае для анализа алгебраической задачи необходимо использовать подходящий итерационный метод, выбор которого чаще всего зависит от постановки исходной задачи.

Например, предположим, что коэффициент λ в уравнении (2.7) зависит от решения: $\lambda = \lambda(x, u)$. В этом случае дискретная модель будет аналогична линейной задаче (2.25), однако элементы матрицы будут зависеть от решения:

$$A(U)U = b. \tag{2.39}$$

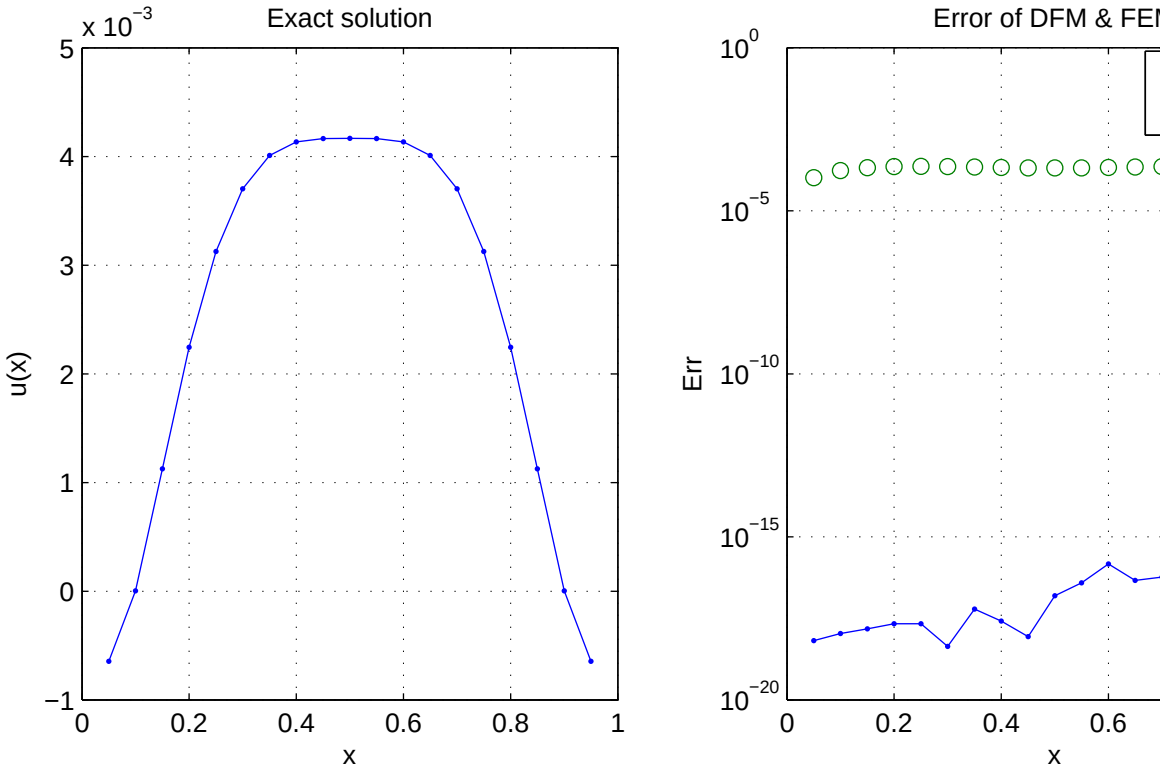

Рис. 2.1. Точное решение задачи (2.7)–(2.8) для следующих значений входных данных: $\lambda(x) = 1$, $q(x) = 0$, $f(x) = 32(x - 0.5)^4 - 4(x - 0.5)^2$ (слева) и погрешности метода конечных разностей и конечных элементов (справа).

Для решения такой нелинейной системы (2.39) во многих практически значимых случаях может быть эффективно использован следующий итерационный метод:

$$A(U^{(k)})U^{(k+1)} = b, \quad k = 1, 2, \ldots, \quad U^{(0)} = 0. \tag{2.40}$$

Возможные проблемы, связанные с отсутствием сходимости итераций (2.40) могут быть устранены с помощью введения итерационного параметра в виде релаксационной процедуры:

$$A(\tilde{U}^{(k)})U^{(k+1)} = b, \quad \tilde{U}^{(k)} = (1 - \sigma)\tilde{U}^{(k-1)} + \sigma U^{(k)}, \quad 0 < \sigma \leqslant 1. \tag{2.41}$$

Подводя итог отметим наиболее примечательные особенности методов конечных разностей и конечных элементов. Прежде всего, на протяжении более половины столетия эти два подхода зарекомендовали себя как наиболее универсальные методы численного анализа дифференциальных краевых задач. Эффективность данных методов во многом связана со структурой получаемых при этом дискретных моделей в виде систем с разреженной матрицей ленточного типа, допускающей эффективное

обращение. Как следствие, методы конечных разностей и конечных элементов во многом отвечают компромиссу между достаточно высокой точностью и умеренной вычислительной сложностью.

Эти методы легко обобщаются на случай неоднородных и адаптивных сеток, что важно при моделировании задач с неоднородными решениями, включая сингулярные режимы и пространственно-локализованные особенности поведения. Наконец, за последние десятилетия разработано большое число модификаций данных методов, которые позволяют улучшить их вычислительные характеристики как в общем случае, так и в случае решения важных частных задач. Достоинства и недостатки этих двух классов методов могут быть кратко резюмированы следующим образом: если вам необходимо решить дифференциальную краевую задачу как можно быстрее и вы не имеете опыта в области численного анализа, тогда разностный метод, пожалуй, может быть лучшим выбором. Однако, если вы ищите мощный вычислительный аппарат, который позволит решить не только текущие проблемы, но способен обеспечить надежную поддержку при развитии и совершенствовании дифференциальной модели, тогда стоит обратить внимание на метод конечных элементов. Преимущества метода конечных элементов проявляется в еще большей степени при решении дифференциальных задач с частными производными.

Упражнение 2.1.

1. Оцените порядок аппроксимации разностной схемы для случая задачи, рассмотренной в примере выше когда правая часть уравнения аппроксимируется следующим образом

$$b_n \simeq f(x_n) + \left(f(x_{n-1}) - 2f(x_n) + f(x_{n+1})\right)/12.$$

2. Проверьте ответ на вопрос упражнения 1, используя соответствующие изменения в программной реализации алгоритма для оценки фактического порядка точности метода. Сравните оценки точности и порядка аппроксимации схемы.

3. Вычислите значения компонент матрицы $\{A_{km}\}$ для метода конечных элементов (2.22)–(2.23) для случая краевой задачи (2.7)–(2.8) с коэффициентами $\lambda(x) = 1+x^2$ и $q(x) = 0$. Какие квадратурные формулы следует использовать, чтобы полученная в итоге схема конечных элементов была эквивалентна разностной схеме (2.14)?

4. Сравните время решения задачи с использованием метода конечных разностей и конечных элементов в рассмотренном выше примере, используя функции Matlab **tic** и **toc**. Какой из методов представляется более эффективным?

5. Для решения краевой задачи (2.7)–(2.8) выполняется следующее интегральное тождество:

$$\int_0^1 f\,dx = \lambda(x)\,\frac{du(x)}{dx}\bigg|_{x=1} - \lambda(x)\,\frac{du(x)}{dx}\bigg|_{x=0}. \tag{2.42}$$

Выведите дискретный аналог данного тождества для разностной схемы (2.15). Дискретный аналог подразумевает использование конечных разностей и квадратурных формул вместо соответствующих производных и интегралов в исходной непрерывной модели. Проверьте выполнение полученного тождества на основе численных

экспериментов с использованием соответствующих дополнений в программную реализацию рассмотренного выше примера.

6. Предложите разностную аппроксимацию краевых условий Неймана:

$$\lambda(x)\frac{du(x)}{dx}\,|_{x=0} = \mu u(x)\,|_{x=0}\,, \quad \lambda(x)\frac{du(x)}{dx}\,|_{x=0} = \mu. \tag{2.43}$$

Одно из возможных решений состоит в использовании уравнения (2.9).

7. Постройте конечно-разностный и конечно-элементный методы для численного решения задачи:

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{du}{dr}\right) = f(r), \quad x \in (0,1), \tag{2.44}$$

$$\left.\frac{du}{dr}\right|_{r=0} = 0, \quad u(1) = 0. \tag{2.45}$$

Используйте дискретизацию задачи на сетке с фиктивным узлом в точке $r = -h/2$, где h — размер шага сетки.

2.3. Спектральные методы

Идея спектральных методов состоит в том, что искомое приближенное решение представляется в виде линейной комбинации некоторого множества базисных бесконечно дифференцируемых функций $\psi_n(x)$, $n = 0, 1, 2, \ldots, N-1$:

$$u(x) \approx \tilde{u}(x) = \sum_{n=0}^{N-1} a(n)\psi_n(x), \quad x \in [a,b], \tag{2.46}$$

где коэффициенты $a(n)$ определяются таким образом, что разность между искомой функция $u(x)$ и ее приближенным представлением $\tilde{u}(x)$ была минимальной в определенном смысле. В отличие от метода конечных элементов, для спектральных методов не требуется ограниченности носителя базисных функций (сравните представления (2.46) и (2.6)).

Для **спектральных методов коллокации** искомое решение представляется в виде:

$$u(x) \approx \tilde{u}(x) = \sum_{n=0}^{N-1} \frac{\alpha(x)}{\alpha(x_n)}\hat{u}_n\phi_n(x), \quad x \in [a,b] \tag{2.47}$$

где $\alpha(x)$ — некоторая весовая функция, $u(x_n) = \tilde{u}(x_n) = \hat{u}_n$, а узлы интерполяции x_n, $n = 0, 1, \ldots, N-1$ называются **точками коллокации**: $a \leqslant x_0 < x_1 < \cdots < x_N \leqslant b$. В качестве множества базисных функций $\phi_n(x)$ наиболее предпочтительным представляется использование систем ортогональных алгебраических или тригонометрических полиномов, $(\phi_n(x), \phi_k(x)) = \delta_{nk}$, где $(\cdot,\cdot)$ и δ_{nk} — соответственно скалярное произведение и символ Кронекера:

$$(\phi_n(x), \phi_k(x)) = \int_a^b \eta(x)\phi_n(x)\phi_k(x)\,dx, \quad \delta_{nk} = \begin{cases} 1, & n = k, \\ 0, & n \neq k. \end{cases} \tag{2.48}$$

Здесь $\eta(x)$ — некоторая весовая функция. В случае ортогонального базиса коэффициенты $a(n)$ в разложении (2.46) вычисляются следующим образом

$$a(n) = (u(x), \phi_n(x)) \,. \tag{2.49}$$

Для вычисления интегралов в скалярном произведении (2.49) могут быть использованы квадратурные формулы, определенные на множестве узлов сетки $x_k \in [a, b]$. Спектральные методы, использующие дискретное представление функций принято называть **псевдоспектральными**.

Типичным примером спектральных методов является **метод Фурье**, основанный на использовании тригонометрических базисных функций, периодических на некотором интервале $x \in [0, L]$:

$$\psi_n(x_k) = \exp\left(\frac{i2\pi n}{L} x_k\right), \quad x_k = kh, \quad k = 0, 1, 2, \ldots N-1, \quad h = L/N. \tag{2.50}$$

Отметим, что переход от сеточной функции $\tilde{u}(x_k)$ к коэффициентам Фурье $a(n)$ и обратно,

$$a(n) = \sum_{k=0}^{N-1} \tilde{u}(x_k) \exp\left(-\frac{i2\pi n}{L} x_k\right), \quad n = 0, 1, \ldots N-1, \tag{2.51}$$

$$\tilde{u}(x_k) = \frac{1}{N} \sum_{n=0}^{N-1} a(n) \exp\left(\frac{i2\pi n}{L} x_k\right), \qquad k = 0, 1, \ldots N-1, \tag{2.52}$$

известен как **прямое** и **обратное дискретное преобразование Фурье**. Данное преобразование может быть представлено как произведения матрицы преобразования Фурье на соответствующий вектор:

$$\mathbf{a} = F\mathbf{u}, \quad \mathbf{u} = F^{-1}\mathbf{a}, \tag{2.53}$$

где

$$F = \frac{1}{\sqrt{N}} \begin{bmatrix} \psi_{0,0} & \psi_{0,1} & \cdots & \psi_{0,N-1} \\ \psi_{1,0} & \psi_{1,1} & \cdots & \psi_{1,N-1} \\ \cdots & \cdots & \cdots & \cdots \\ \psi_{N-1,0} & \psi_{N-1,1} & \cdots & \psi_{N-1,N-1} \end{bmatrix}, \tag{2.54}$$

$$F^{-1} = \frac{1}{\sqrt{N}} \begin{bmatrix} \psi^{-1}_{0,0} & \psi^{-1}_{0,1} & \cdots & \psi^{-1}_{0,N-1} \\ \psi^{-1}_{1,0} & \psi^{-1}_{1,1} & \cdots & \psi^{-1}_{1,N-1} \\ \cdots & \cdots & \cdots & \cdots \\ \psi^{-1}_{N-1,0} & \psi^{-1}_{N-1,1} & \cdots & \psi^{-1}_{N-1,N-1} \end{bmatrix}, \tag{2.55}$$

$$\mathbf{u} = (\tilde{u}(x_0), \tilde{u}(x_1), \ldots, \tilde{u}(x_{N-1}))^T, \quad \mathbf{a} = (a(0), a(1), \ldots, a(N-1))^T, \tag{2.56}$$

$$\psi_{km} = \exp(-i2\pi km/N). \tag{2.57}$$

Постоянные $\omega_k = 2\pi k/L$ имеют смысл частот соответствующих базисных функций. Матрицы F и $F^{-1} = F^*$ — матрицы прямого и обратного дискретных преобразований Фурье. Вычисление прямого и обратного преобразований Фурье согласно (2.53) имеет вычислительную сложность $O(N^2)$.

Одно из замечательных свойств метода Фурье состоит в том, что дискретное преобразование Фурье моет быть реализовано с вычислительными затратами порядка $O(N\log(N))$ при использовании **алгоритма быстрого преобразования Фурье БПФ** . В случае когда N равно целой степени двойки, $N = 2^m$, $m = 1, 2, 3, \ldots$, алгоритм БПФ может быть представлен в виде особого рода факторизации **циркулянтной матрицы**, каковой является матрица преобразования Фурье [1]:

$$F = S_m S_{m-1}, \cdots S_1, \tag{2.58}$$

где S_k, $k = 1, 2, \ldots m$ — разреженные матрицы с не более чем двумя ненулевыми элементами в каждой строке. Как следствие, вычислительная сложность умножения $S_k\mathbf{u}$ имеет порядок $O(N)$. В итоге вычислительная сложность цепочки произведений $S_m S_{m-1}, \cdots S_1\mathbf{u}$ составляет $O(Nm) = O(N\log N)$.

Коэффициенты Фурье $a(k)$, равно как и преобразуемый вектор $u(x)$, полагаются периодическими:

$$u(x) = u(x + L), \quad a(k) = a(k + sN), \quad s = \pm 1, \pm 2, \ldots. \tag{2.59}$$

С учетом периодичности можно использовать произвольное множество из N последовательных значений в дискретном представлении функции $u(x_k)$ и ее коэффициентов Фурье $a(k)$, $k = m, m+1, \ldots, m+N-1$. Для многих приложений используется симметричная частотная полоса:

$$\tilde{u}(x_k) = \frac{1}{N}\sum_{n=-N/2}^{N/2-1} a(n)\exp\left(\frac{i2\pi n}{L}x_k\right), \qquad k = 0, 1, \ldots N-1. \tag{2.60}$$

Приближенное представление периодической функции (2.60) может рассматриваться как конечный отрезок ряда Фурье. Если функция $u(x)$ является достаточно гладкой (имеет p непрерывных производных) тогда коэффициенты Фурье быстро убывают

$$a(n) = O(n^{-(p+2)}). \tag{2.61}$$

Погрешность метода Фурье ассоциируется с **эффектом маскировки частот**, когда высокочастотные компоненты $a(k)$ за пределами частоты Найквиста, $|k| > N/2$, присоединяется к компоненте $a(m)$ согласно соотношению (2.59),

[1] Специальный класс Теплицевых матриц в которых каждая следующая строка характеризуется вращением на один элемент вправо по отношению к предыдущей строке. Замечательные свойства циркулянтных матриц в численном анализе с связаны с возможностью приведения их к диагональному виду с помощью алгоритма быстрого дискретного преобразования Фурье

$|m| = |k| - N[k/N]$. Здесь $[k/N]$ означает целую часть отношения k/N. Для получения приемлемой точности частота дискретизации должна выбираться таким образом, чтобы коэффициенты Фурье за пределами частоты Найквиста были пренебрежимо малы. Это может быть обеспечено выбором достаточно большого N.

Дискретное преобразование Фурье с использованием алгоритма БПФ можно примяться для численного дифференцирования. Действительно, как следует из уравнения (2.60)

$$\frac{d^m\tilde{u}(x)}{dx^m} = \frac{1}{N}\sum_{n=-N/2}^{N/2-1}\left(\frac{i2\pi n}{L}\right)^m a(n)\exp\left(\frac{i2\pi n}{L}x\right). \tag{2.62}$$

Таким образом, численное дифференцирование с использованием преобразования Фурье сводится к поэлементному умножению коэффициентов Фурье $a(n)$ на соответствующий вектор частот $i\omega_n$, $n = -N/2, -N/2-1, \ldots N/2-1$.

Используя разложение (2.47) оператор дифференцирования функции, заданной вектором значений на равномерной сетке, может быть представлен в виде матрицы $D^{(m)}$ размерности $N \times N$, которая называется **спектральной матрицей дифференцирования**:

$$D^{(m)}_{k,n} = \frac{d^m}{dx^m}\left[\frac{\alpha(x)}{\alpha(x_n)}\phi_n(x)\right]_{x=x_k} \tag{2.63}$$

Процедура численного дифференцирования сводится к умножению спектральной матрицы дифференцирования на соответствующую сеточную функцию.

$$\mathbf{u}^{(m)} = D^{(m)}\mathbf{u}, \tag{2.64}$$

где $\mathbf{u} = (\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_{N-1})^T$, $\tilde{u}_k = \tilde{u}(x_k)$, $\mathbf{u}^{(m)} = (\tilde{u}_0^{(m)}, \tilde{u}_1^{(m)}, \ldots, \tilde{u}_{N-1}^{(m)})^T$, $u_k^{(m)} = \left[\dfrac{d^m\tilde{u}}{dx^m}\right]_{x=x_k}$.

Матрица спектрального дифференцирования произвольного порядка m в методе Фурье может быть вычислена следующим образом:

$$D^{(m)} = F^{-1}\Omega^m F, \tag{2.65}$$

где F и F^{-1} — матрицы прямого и обратного преобразования Фурье, Ω — диагональная матрица $N \times N$, такая, что $\Omega_{k,k} = i\omega_k$, $\Omega_{k,m\neq k} = 0$.

В отличие от методов конечных разностей и конечных элементов, где матрица оператора дифференцирования является разреженной ленточной, матрица спектрального дифференцирования является полной. Как следствие, вычислительная сложность спектральных методов существенно выше по сравнению с разностными и конечно элементными. Тем не менее, высокая вычислительная сложность спектральных моделей в случае достаточно гладких решений компенсируется высокой точностью, что обеспечивает в итоге преимущества данного класса методов. Кроме того, спектральные методы с матрицей дифференцирования теплицева типа при использовании БПФ алгоритма способны обеспечить непревзойденную эффективность.

Пример 2.2. Рассмотрим пример спектрального дифференцирования с использованием метода Фурье, применительно к функциям, имеющим различную степень гладкости:

$$u_1(x) = \exp(-16x^4), \quad u_2(x) = \begin{cases} \exp(-16x^2), & x \in [-\pi, 0], \\ \exp(-16x^4), & x \in [0, \pi]. \end{cases} \tag{2.66}$$

Отметим, что функция $u_1(x)$ является бесконечно дифференцируемой, в то время как функция $u_2(x)$ имеет лишь одну непрерывную производную. Результаты численных экспериментов и программная реализация спектрального дифференцирования Фурье представлены ниже.

```
% Погрешность спектрального метода дифференцирования Фурье
%  в зависимости от гладкости дифференцируемой функции
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
N = 128/1;
h = 2*pi/N;
x = -pi:h:pi-h; %Note, the last point is omitted
d = 1i*fftshift(-N/2:N/2-1);
%%%% гладкая функция u_1(x)%%%%%%%%%%%%%%%%%%%%%%%
u1 = exp(-8*x.^4);
%%%% не гладкая функция u_1(x)%%%%%%%%%%%%%%%%%%%
u2(1:N/2) = exp(-8*x(1:N/2).^4);
u2(N/2+1:N) = exp(-8*x(N/2+1:N).^2);
%%% точные производные %%%%%%%%%%%%%%%%%%%%%%%%%%
du1 = -32*x.^3.*exp(-8*x.^4);
du2(1:N/2) = -32*x(1:N/2).^3.*exp(-8*x(1:N/2).^4);
du2(N/2+1:N) = -16*x(N/2+1:N).*exp(-8*x(N/2+1:N).^2);
%%% Спектральные производные %%%%%%%%%%%%%%%%%%%%%%
Du1 = ifft(d.*fft(u1));
Du2 = ifft(d.*fft(u2));
%%%%error of the spectral differentiation %%%%%%
err1 = abs(Du1-du1);
err2 = abs(Du2-du2);
m=1:4:length(x);
semilogy(x(m),err1(m),'o-',x(m),err2(m),'.-','LineWidth',1)
legend('smooth','non smooth')
xlabel('x')
ylabel('Погрешность спектрального дифференцирования')
grid
```

Представленные на рис. 2.2 результаты показывают существенную зависимость погрешности спектрального дифференцирования Фурье от гладкости дифференцируемой функции. Для бесконечно дифференцируемой функции $u_1(x)$ погрешность спектрального дифференцирования достигает порога вычислительной погрешности на сравнительно грубой сетке $N \geqslant 128$. В случае функции $u_2(x)$ производные высшего порядка разрывные при $x = 0$. Как следствие, ошибка спектрального дифференцирования в последнем случае существенно выше и медленно убывает с уменьшением шага сетки.

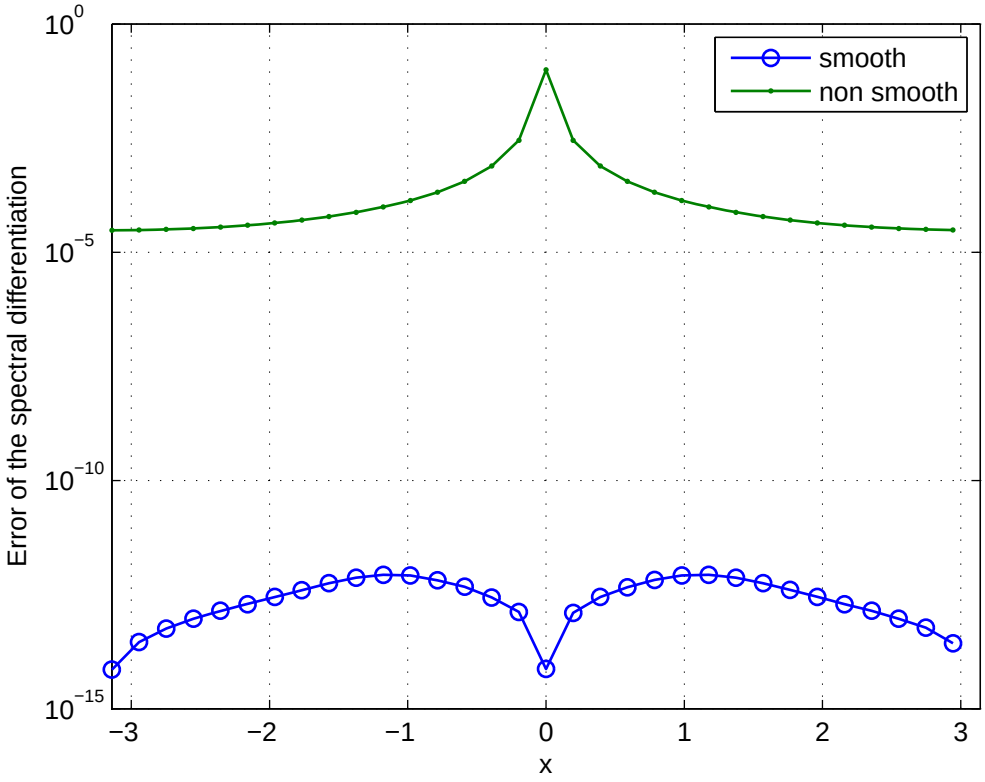

Рис. 2.2. Погрешности спектрального дифференцирования Фурье в случае гладкой и не гладкой функции при $N = 128$.

Рассмотренный выше пример показывает высокую эффективность спектрального метода в случае достаточной гладкости искомого решения. Однако, если функция или ее производные разрывны, то спектральные методы теряют свои преимущества. Отметим, что в спектральном методе Фурье требования гладкости распространяются также на периодическое продолжение функции за пределы рассматриваемого интервала. В силу последнего, метод Фурье может работать только с периодическими краевыми условиями или с финитными функциями, обращающимися в ноль вне некоторого интервала. Для нулевых краевых условий наиболее подходящим представляется спектральный метод, использующий синус преобразование Фурье.

Наряду с методом Фурье, один из наиболее эффективных методов решения дифференциальных задач является **спектральный метод Чебышева**, который может использоваться для произвольных краевых условий.

Матрица спектрального дифференцирования Чебышева вычисляется на основе интерполяционного представления (2.47) с весовой функцией $\alpha(x) \equiv 1$

и интерполяционными узлами, определяемыми точками экстремумов полинома Чебышева первого рода:

$$T_n(x) = \cos\left(n \arccos(x)\right), \quad x \in [-1, 1], \quad n = N - 1 :$$

$$x_k = -\cos\frac{(k-1)\pi}{N-1}, \quad k = 1, ..., N \tag{2.67}$$

Базисные функции определяются следующим образом:

$$\phi_k(x) = \frac{(-1)^k}{c_k}\frac{1-x^2}{(N-1)^2}\frac{T'_N(x)}{x - x_k}, \quad c_1 = c_N = 2, \quad c_2 = ... = c_{N-1} = 1. \tag{2.68}$$

Согласно (2.50), (2.68) элементы матрицы спектрального дифференцирования Чебышева вычисляются по следующим формулам:

$$D^{(1)}_{1,1} = \frac{2(N-1)^2 + 1}{6}, \tag{2.69}$$

$$D^{(1)}_{N,N} = -\frac{2(N-1)^2 + 1}{6}, \tag{2.70}$$

$$D^{(1)}_{k,k} = \frac{-x_k}{2(1-x_k^2)}, \quad k = 2, ..., N-1, \tag{2.71}$$

$$D^{(1)}_{k,n} = \frac{c_k}{c_n}\frac{(-1)^{k+n}}{(x_k - x_n)}, \quad k \neq n, k, n = 1, ..., N, \tag{2.72}$$

Фактически каждый элемент матрицы дифференцирования $D^{(1)}_{k,n}$ численно равен производной интерполяционного полинома $\phi_n(x)$ (2.68) в точке $x = x_k$. Для вычисления производных более высокого порядка с помощью матрицы спектрального дифференцирования Чебышева может быть использовано следующее тождество:

$$D^{(m)} = {D^{(1)}}^m. \tag{2.73}$$

Отметим, что для некоторого вида матриц дифференцирования формула (2.73), вообще говоря, не вполне подходит для вычисления производных высшего порядка. Например, матрица разностного дифференцирования $D^{(1)}_C$ имеет вид

$$D^{(1)}_C = \frac{1}{2h}\begin{bmatrix} 0 & 1 & 0 & & 0 \\ -1 & 0 & \ddots & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & \ddots & \ddots & 0 & 1 \\ 0 & & 0 & -1 & 0 \end{bmatrix}. \tag{2.74}$$

Однако, для аппроксимации второй производной более предпочтительным представляется использовать произведение матриц правой и левой разностных производных вместо квадрата матрицы центральной производной:

$$D^{(2)} = D^{(1)}_B D^{(1)}_F = D^{(1)}_F D^{(1)}_B, \tag{2.75}$$

где

$$D_B^{(1)} = \frac{1}{h}\begin{bmatrix} 1 & 0 & 0 & & 0 \\ -1 & 1 & \ddots & \ddots & \\ 0 & -1 & \ddots & \ddots & 0 \\ & \ddots & \ddots & 1 & 0 \\ 0 & & 0 & -1 & 1 \end{bmatrix} \quad D_F^{(1)} = \frac{1}{h}\begin{bmatrix} -1 & 1 & 0 & & 0 \\ 0 & -1 & \ddots & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & \ddots & \ddots & -1 & 1 \\ 0 & & 0 & 0 & -1 \end{bmatrix}. \quad (2.76)$$

Используя понятие матрицы дифференцирования мы получаем вполне прозрачное понимание общих принципов построения спектральных численных методов. В частности, построение спектральной дискретной модели состоит в замене дифференциального бесконечномерного оператора на спектральную матрицу дифференцирования, в которой учтены краевые условия.

В случае метода Фурье краевые условия полагаются периодическими, что автоматически учитывается выбором периодических базисных функций. Для спектрального метода Чебышева краевые условия произвольного вида могут быть заданы с помощью соответствующей модификации крайних строк матрицы дифференцирования. Рассмотрим простой пример, демонстрирующий постановку краевых условий.

Пример 2.3. Построим спектральный метод Чебышева для уравнения второго порядка с краевыми условиями Дирихле и исследуем его эффективность на задачах с различной степенью гладкости входных данных:

$$\frac{d^2u}{dx^2} = f(x), \quad x \in [-1, 1], \quad (2.77)$$

$$u(-1) = a, \quad u(1) = b. \quad (2.78)$$

В первом случае полагаем

$$f(x) = -x^2/12, \quad (2.79)$$

и решение задачи является гладким:

$$u(x) = x^4 - 1. \quad (2.80)$$

Во втором случае функция правой части полагается разрывной:

$$f(x) = 50\,\mathrm{sign(x)}. \quad (2.81)$$

Как следствие, производные решения задачи второго и более высокого порядка также будут разрывными. Точное решение имеет вид

$$u(x) = 25\,\mathrm{sign(x)}\,\left(\mathrm{x}^2 - |\mathrm{x}|\right). \quad (2.82)$$

Дискретизация области определения решения и замена дифференциального оператора соответствующей матрицей дифференцирования приводит к следующей системе дифференциальных уравнений

$$D^{(2)}U = F, \quad (2.83)$$

где $F = (a, f(x_1), f(x_2), \ldots, f(x_{N-2}), b)^T$. Отметим, что первую и последнюю строки матрицы дифференцирования следует модифицировать согласно краевым условиям, чтобы решение алгебраической задачи (2.83) было определено однозначно. В случае нулевых краевых условий эти строки (равно как первый и последний столбцы матрицы $D^{(2)}$) могут быть просто удалены. В соответствии с этим удаляются первый и последний элементы вектора F.

Для краевых условий Дирихле общего вида элементы первой и последней строк полагаются равными нулю, за исключением диагональных элементов, значения которых задаются равными единице. При этом первый и последний элемент вектора правой части F задаются равными соответствующим краевым значения.

В случае краевых условий Неймана, например,

$$\left(\frac{du}{dx} - \alpha u\right)\Big|_{x=-1} = \beta, \tag{2.84}$$

первая строка матрицы дифференцирования $D^{(2)}$, отвечающая граничной точке $x = -1$, заменяется соответствующей строкой матрицы $D^{(1)}$ с добавлением к диагональному элементу этой строки слагаемого $-\alpha$. Первому элементу вектора F соответственно присваивается значение β.

```
%% Решение уравнения u"=sign(x)*50 %%%%
%%% с разрывной правой частью %%%%%%%%
%% Конечные разности в сравнении со спектральным методом %%
clear; k = 0;
for H = [1e-1,5e-2,2.5e-2,1e-2,5e-3,2e-3,1e-3]
k = k+1;
h = H;
x = -1+h:h:1-h;
N = length(x);
e = ones(N,1)/h^2;
%%%% конечноразностная матрица дифференцирования %%%%%%%%%%%%%
A = spdiags([e,-2*e,e],-1:1,N,N);
f = -12*x.^2;
u = A\(f(:));
%%%% = точное решение =%%%%%%%%%%%%%%%%%%%%%%
U = -x.^4+1;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
j = 0:N-1;
xx = -cos(j.*pi/(N-1));
%%%% матрица спектрального дифференцирования Чебышева %%%%%%%%%%
D = gallery('chebspec',N);
AA = D*D; %вторая производная
%%%% = Краевые Условия = %%%%%%%%%%%%%%%%%%
AA(1,1) = 1;
AA(1,2:end) = 0;
AA(N,N) = 1;
AA(N,1:end-1)=AA(N,1:end-1)*0;
```

```
f = -12*xx.^2;
f(end) = 0;
f(1) = 0;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
uu = AA \ f(:);
%%%% = точное решение =%%%%%%%%%%%%%%%%%%%%%%%
UU=-xx.^4+1;;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
err_fd(k) = norm(U(:)-u)/norm(U);
err_sc(k) = norm(UU(:)-uu)/norm(UU);
Ns(k) = N;
end
loglog(Ns,err_fd,'o-',Ns,err_sc,'.-','LineWidth',2)
legend('FD','SC')
title('Err. Finite-Difference vs. Spectral Chebyshev')
xlabel('N');
ylabel('||\delta||/||u||');
axis([15 3000 1e-17 1]);
grid
figure
plot(xx,UU,'r','LineWidth',2)
title('Solution of the Equation u"=sign(x)*50')
xlabel('x');
ylabel('u(x)');
axis([-1 1 0 1.5]);
grid
```

Результаты, представленные на рис. 2.3 показывают, что спектральный метод Чебышева демонстрирует очень высокую точность на относительно грубой сетке. Матрица спектрального дифференцирования Чебышева дает практически точные результаты при дифференцировании полиномиальной функции степени $(N-1)$. В этом случае погрешность ограничена только вычислительной погрешностью (погрешностью округления чисел с плавающей запятой).

Метод конечных разностей в данном случае имеет второй порядок точности. Для достижения точности, сравнимой с точностью спектрального метода, разностный метод требует очень подробную сетку с числом узлов порядка $N \simeq 10^5$.

Тем не менее, преимущества спектрального метода Чебышева могут проявляться только в случае достаточно гладкого решения. При численном решении задачи (2.77)–(2.78), (2.80) ситуация коренным образом меняется. Точное решение имеет вид кусочно квадратичной функции (2.82). Локальная погрешность метода конечных разностей определяется производной четвертого порядка от решения (см. (2.10)). Как следствие, на параболическом решении ошибка дискретизации разностной схемы обращается в ноль. Данный факт подтверждается результатами численного эксперимента, представленными на рис. 2.3.

Что касается спектрального метода, то в случае негладкого решения его погрешность достаточно велика и сравнима с погрешностью разностного метода при

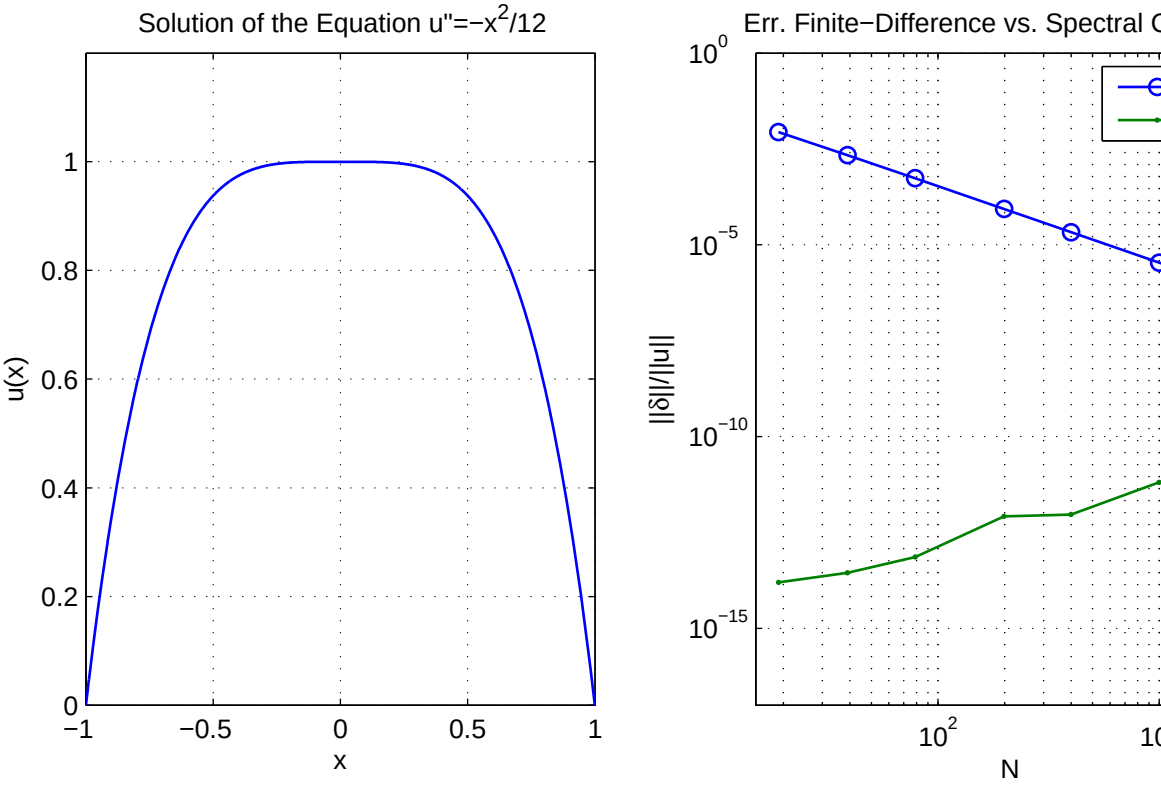

Рис. 2.3. Решение задачи (2.77)–(2.79) (слева) и динамика относительной ошибки спектрального метода Чебышева (SC) и метода конечных разностей в зависимости от числа узлов сетки (справа)

решении задачи (2.77)–(2.79) (сравните результаты, представленные на рис. 2.3 и рис. 2.3).

Рассмотренный пример показывает, что точность спектрального метода (в данном случае это метод Чебышева) существенно зависит от гладкости входных данных задачи. Как правило, погрешность метода имеет степенную скорость сходимости:

$$\|\delta\| = \|U - u\| \leqslant CN^{-m}, \tag{2.85}$$

где C — некоторая постоянная, N — число узлов сетки, m характеризует скорость сходимости. Для большинства известных разностных методов и методов конечных элементов $m = 2$ или $m = 4$. Скорость сходимости спектральных методов существенно выше и в большинстве случаев ограничена не столько внутренними особенностями метода, сколько гладкостью решения. Для бесконечно дифференцируемых решений оценка (2.85) может выполняться для произвольных $m > 0$. Для решений, имеющих аналитическое продолжение на комплексную плоскость сходимость спек-

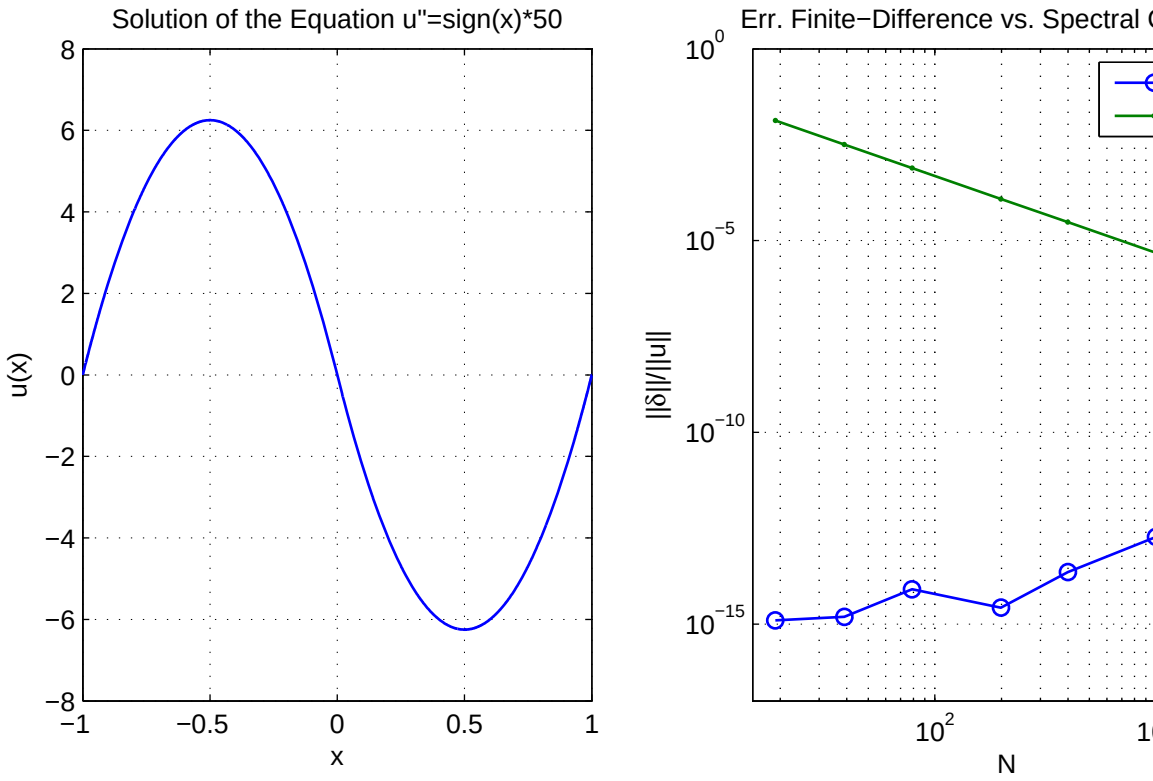

Рис. 2.4. Решение задачи (2.77)–(2.78), (2.80) (слева) и динамика погрешности спектрального метода Чебышева (SC) и метода конечных разностей в зависимости от числа узлов сетки (справа).

тральных методов имеет экспоненциальный характер, т.е. быстрее, чем степенная сходимость (2.85):

$$\|\delta\| = \|U - u\| \leqslant \exp(-CN). \tag{2.86}$$

В случае функций, производные которых вплоть до порядка $p-1$, принадлежат классу интегрируемых с квадратом, скорость сходимости ν-й производной, $\nu < p$, имеет следующую оценку:

$$\|u^{(\nu)} - D^{(\nu)}U\| \leqslant CN^{-(p-\nu+1)}, \tag{2.87}$$

для $p \geqslant 1$.

В случае задачи (2.77)–(2.78), (2.80), с учетом того, что вторая производная задана уравнением (2.81) и $p = 3$, согласно оценке (2.87) мы имеем второй порядок скорости сходимости, что подтверждается результатами численных экспериментов, представленных на рис. 2.3.

Для большинства практически значимых случаев оценка гладкости решения представляет собой нетривиальную задачу. В силу этого наиболее простой способ убедиться в сходимости спектрального метода состоит в сравнении приближенных

решений, полученных на последовательности 2-3 сеток с различным числом узлов. Кроме того, существует несколько эмпирических правил, позволяющих оценить адекватный размер шага для получения требуемую точность. Например, при моделировании колебательных процессов разумно соизмерять выбор размера шага с периодом колебаний или длиной волны. В частности, наблюдения показывают, что для получения относительной погрешности 5% при моделировании волновых процессов со средней длиной волны λ необходимое пространственное разрешение сетки для методов конечных разностей второго порядка точности может быть обеспечено при размере шага дискретизации $h \simeq \lambda/20$, в то время как для спектральных методов Фурье (Чебышева) $h \simeq \lambda/3.5$. Для достижения относительной погрешности 1% соответственно требуется размер шага для разностных методов $h \simeq \lambda/40$, а для спектральных методов по-прежнему $h \simeq \lambda/3.5$ (подробнее см. [8]).

Еще одно эмпирическое правило выбора шага: при моделировании сингулярных решений на отрезке $x \in [-1, 1]$, когда масштаб сингулярности $2\varepsilon << 1$ относительная точность порядка ε обеспечивается спектральным методом Чебышева при размере шага

$$h \simeq \frac{\lambda}{2\log(\varepsilon)}. \tag{2.88}$$

Пример 2.4. Для лучшего понимания преимуществ спектральных методов сравним его эффективность с другими численными методиками, например, с методом сплайн-коллокации 5-го порядка точности, реализованным в стандартной функции MATLAB **bvp5c**.

В качестве тестовой задачи рассмотрим систему дифференциальных уравнений

$$\begin{cases} \dfrac{du}{dx} = ik\exp(i2\Delta x)v, \\ \\ \dfrac{dv}{dx} = ik\exp(-i2\Delta x)u, \end{cases} \quad , \quad x \in [-1, 1], \tag{2.89}$$

где $i = \sqrt{-1}$, с краевыми условиями

$$u(-L) = a,\ v(L) = b. \tag{2.90}$$

Данная задача описывает распространение электромагнитных волн в среде с периодической модуляцией показателя преломления. Для оценки погрешности приближенного решение мы будем использовать аналитическое выражение для коэффициента отражения волны:

$$R = \frac{|v(-1)|}{|u(-1)|} = \frac{\sinh(\alpha L)}{\sqrt{\cosh^2(\alpha L) - \chi^2}} \tag{2.91}$$

где $\alpha = \sqrt{k^2 - \Delta^2}$, $\chi = \dfrac{\Delta}{k}$.

Используя сетку Чебышевских узлов

$$x_m = -\cos\frac{(m-1)\pi}{N-1}, \quad m = 1, 2, \ldots, N, \tag{2.92}$$

построим дискретную модель на основе спектрального метода Чебышева, которая приводит к следующей системе линейных алгебраических уравнений:

$$AY = F, \tag{2.93}$$

где $Y = (U, V)^T$, $U = (u_0, u_2, \dots u_{N-1})$, $V = (v_0, v_2, \dots v_{N-1})$,

$$A = \begin{pmatrix} D^+ & G^+ \\ G^- & D^- \end{pmatrix}, \tag{2.94}$$

$G^{\pm}$ — диагональные матрицы $N \times N$:

$$g^+_{m,m} = ik \exp(i2\Delta x_m), \quad m = 2, 3, \dots, N, \quad g^+_{1,1} = 0, \tag{2.95}$$

$$g^-_{m,m} = ik \exp(-i2\Delta x_m), \quad m = 1, 2, \dots, N-1, \quad g^-_{N,N} = 0, \tag{2.96}$$

$D^{\pm}$ — матрица спектрального дифференцирования Чебышева размерности $N \times N$, в которой первая (последняя) строка модифицирована в соответствии с граничными условиями: $d^+_{1,1} = d^-_{N,N} = 1$, $d^+_{1,m\neq 1} = d^-_{N,m\neq N} = 0$, а вектор правой части системы F с учетом краевых условий имеет вид: $F = (a, 0, \dots, 0, b)^T$.

Программная реализация алгоритма спектрального метода и решение задачи с использованием функции **bvp5c** представлены ниже. Для более реалистичной оценки вычислительных затрат мы повторяли процедуру несколько раз, выбирая минимальное время решения.

```
%% Спектральный метод Чебышева и метод сплайн-коллокации
%%%% Сравнение эффективности методов %%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%     параметры задачи    %%%%%%%%%%%%%%%%%%%%%%%
k = 2;  d = 5;
a = 1;  b = 0;
alp = sqrt(k^2-d^2); Xi = d/k;
R0=sinh(alp*2)/sqrt((cosh(alp*2))^2-Xi^2);
%% некоторые вспомогательные функции. см >>help bvp5c %%%
dydx = @(x,y)[-1i*k*exp( 1i*2*d*x)*y(2);...
1i*k*exp(-1i*2*d*x)*y(1)];
res = @(ya,yb)[ya(1) - a; yb(2)-b ];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
K = 12;
%%%% Задание точности метода для функции bvp5c %%%%%%%%%%%%%
RelTol = logspace(-2,-9,K);
Ns = round(44./RelTol.^(1/4));
for m=1:K
solinit = bvpinit(linspace(-1,1,Ns(m)),[1;0]);
options = bvpset('RelTol',RelTol(m));
%% процедура повторяется 15 раз и выбирается минимальное время  %
for mm=1:15
tic
sol = bvp5c(dydx,res,solinit,options);
Txx(mm) = toc;
```

```
end
R_cm(m) = abs(sol.y(2,1))./a;
T_cm(m) = min(Txx);
Err_cm(m)=abs(R0-R_cm(m))/R0;
end
%% Последовательность сеток спектрального метода %%
Ns=[18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 48, 64];
for m=1:K
N = Ns(m);
%%% Чебышевская сетка и спектральная матрица дифференцирования %%
x = -cos((0:N-1)*pi/(N-1));
C = gallery('chebspec',N);
%% процедура повторяется 15 раз и выбирается минимальное время  %
for mm=1:15
tic
%% 2N x 2N матрица спектральной задачи %
A = zeros(2*N);
G_p = -diag(1i*k*exp( i*2*d*x(:)));
G_m =  diag(1i*k*exp(-i*2*d*x(:)));
A =[ C, G_p; G_m, C];
%%%%%% краевые условия %%%%%%%%%%%%%%%%%%%%%%%%%%
A(1,:) = A(1,:)*0;      A(1,1) = 1;
A(end,:) = A(end,:)*0;  A(end,end) = 1;
F=zeros(2*N,1);         F(1) = a; F(end) = b;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Y = A\F;
Tx(mm) = toc;
end
R_sm(m) = abs(Y(N+1))./a;
T_sm(m) = min(Tx);
Err_sm(m) = abs(R0-R_sm(m))/R0;
end
loglog(T_cm,Err_cm,'.-',T_sm,Err_sm,'o-','Linewidth',1)
title('Spectral Chebyshev vs. 5-th order Spline-Collocation' )
xlabel('Calculation Time [sec]');
ylabel('Relative Accuracy')
legend('bvp5c', 'SChM'); grid; figure; spy(A)
```

Результаты численных экспериментов представлены на рис. 2.5. Сравнение вычислительных затрат для достижения заданной точности показывает, что спектральный метод Чебышева существенно превосходит в эффективности стандартные средства MATLAB для решения двухточечных краевых задач. Сетка с числом узлов $N = 32$ при использовании спектрального метода позволяет достичь предельно малой погрешности, сравнимой с вычислительной погрешностью. При одинаковых требованиях к точности время решения задачи при использовании метода сплайн-коллокации более чем на два порядка превосходит данный показатель для спектрального метода.

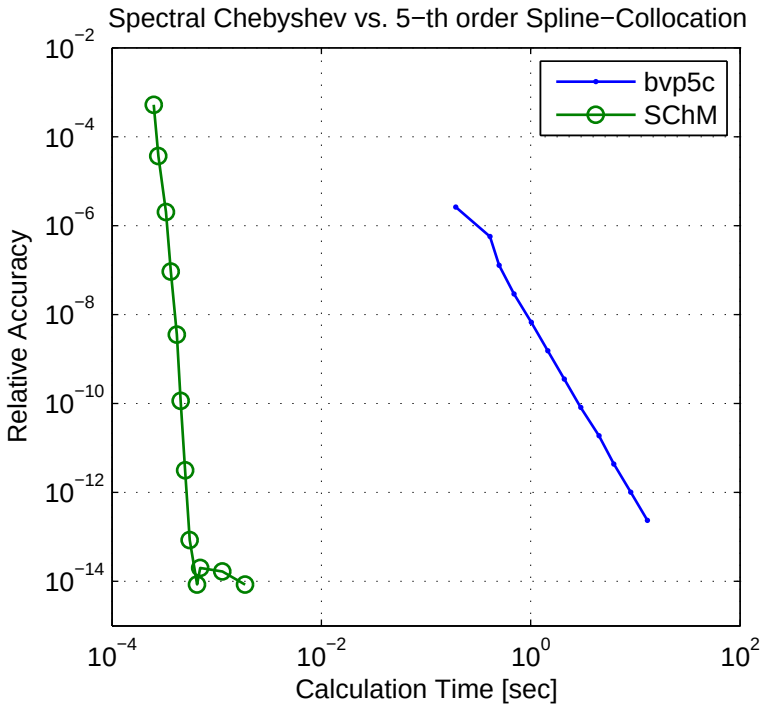

Рис. 2.5. Время решения задачи для достижения заданной точности при использовании спектрального метода Чебышева (SChM) и метода сплай-коллокации 5-го порядка точности, реализованного в функции **bvp5c.m**.

Упражнение 2.2. 1. Воспроизведите пример, представленный выше и убедитесь, что решение задачи (2.89)–(2.90) имеет волнообразный вид и для заданных параметров $\lambda \simeq 2/3$. На примере рассмотренной задачи проверьте выполнения эмпирического правила выбора шага спектральных методов для достижения относительной погрешности 1-4%. Среднее значение шага считать равным $2/N$, где N — число узлов сетки.

2. Проверьте справедливость эмпирического правила (2.88) на примере спектрального дифференцирования функции

$$f(x) = \frac{\lambda^2}{\lambda^2 + x^2}, \quad x \in [-1, 1], \quad \lambda = 0.1, 0.05, 0.01, \tag{2.97}$$

используя матрицу спектрального дифференцирования Чебышева.

3. Для решения задачи (2.89)–(2.90) выполняется следующий закон сохранения:

$$\left[|u|^2 - |v|^2\right] = \text{const}. \tag{2.98}$$

Проверьте выполнение данного закона сохранения для спектрального метода и метода сплай-коллокации, используя программную реализацию методов, рассмотренную в примере выше.

4. Как можно объяснить рост ошибки разностного метода при уменьшении шага сетки в рассмотренном выше примере (см., рис. 2.4).

5. Какое минимальное число узлов сетки требуется для достижения предельной (спектральной) точности при дифференцировании полинома 6-го порядка с использованием матрицы спектрального дифференцирования Чебышева. Проверьте ответ с помощью прямых вычислений, приготовив соответствующий пример произвольного полинома 6-го порядка.

Библиографические ссылки

1. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы. М. : БИНОМ. Лаб. знаний, 2003. 636 с.
2. Норри Д., де Фриз Ж. Введение в метод конечных элементов. - М.: Мир, 1981. - 155 с.
3. Ортега Дж., Пул У. Введение в численные методы решения дифференциальных уравнений. М.: Наука, 1986. - 288 с.
4. Самарский А.А., Гулин А.В. Численные методы. М.: Наука, 1989. 432 с.
5. Калиткин Н. Н. Численные методы. 2 изд. – БХВ-Петербург, 2011, 592 с.
6. Ascher U.M., and Petzold L.R.: *Computer methods for ordinary differential equations and differential-algebraic equations*, Vol. 61. SIAM, Philadelphia (1998).
7. Borse G.J.: *Numerical methods with MATLAB: A resource for scientists and engineers*, International Thomson Publishing, Boston (1996).
8. Boyd J.P.: *Chebyshev and Fourier spectral methods*, Courier Corporation , New York (2001).
9. Butcher J.C.: *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*, Wiley-Interscience, New York, (1987).
10. Fornberg B.: *A Practical Guide to Pseudospectral Methods*. Cambridge University Press, Cambridge (1998).
11. Gockenbac M.S.h: *Understanding and implementing the finite element method*, SIAM, Philadelphia (2006).
12. Golub G.H. and Ortega J.M., *Scientific Computing and Differential Equations: an Introduction to Numerical Methods*, Academic Press inc., London (1991).
13. Gottlieb D. and Orszag S.: The Numerical Analysis of Spectral Methods, SIAM, Philadelphia (1987).
14. Hairer E., Norsett S.P., Wanner G.: *Solving ordinary differential equations. I. Nonstiff problems. Second edition*, Springer Series in Computational Mathematics, 8. Springer-Verlag, Berlin (1993).
15. Hairer E., Wanner G.: *Solving ordinary differential equations. II. Stiff and differential-algebraic problems. Second revised edition, paperback*. Springer Series in Computational Mathematics, 14. Springer-Verlag, Berlin, (2010).
16. Iserles A.: A first course in the numerical analysis of differential equations. Vol. 44. Cambridge University Press, Cambridge (2009).
17. Liu G.R. and Quek S.S.: *The finite element method: A practical course*. Butterworth-Heinemann (2013).
18. Press W.H.: *Numerical recipes. 3rd edition: The art of scientific computing*, Cambridge university press, Cambridge (2007).